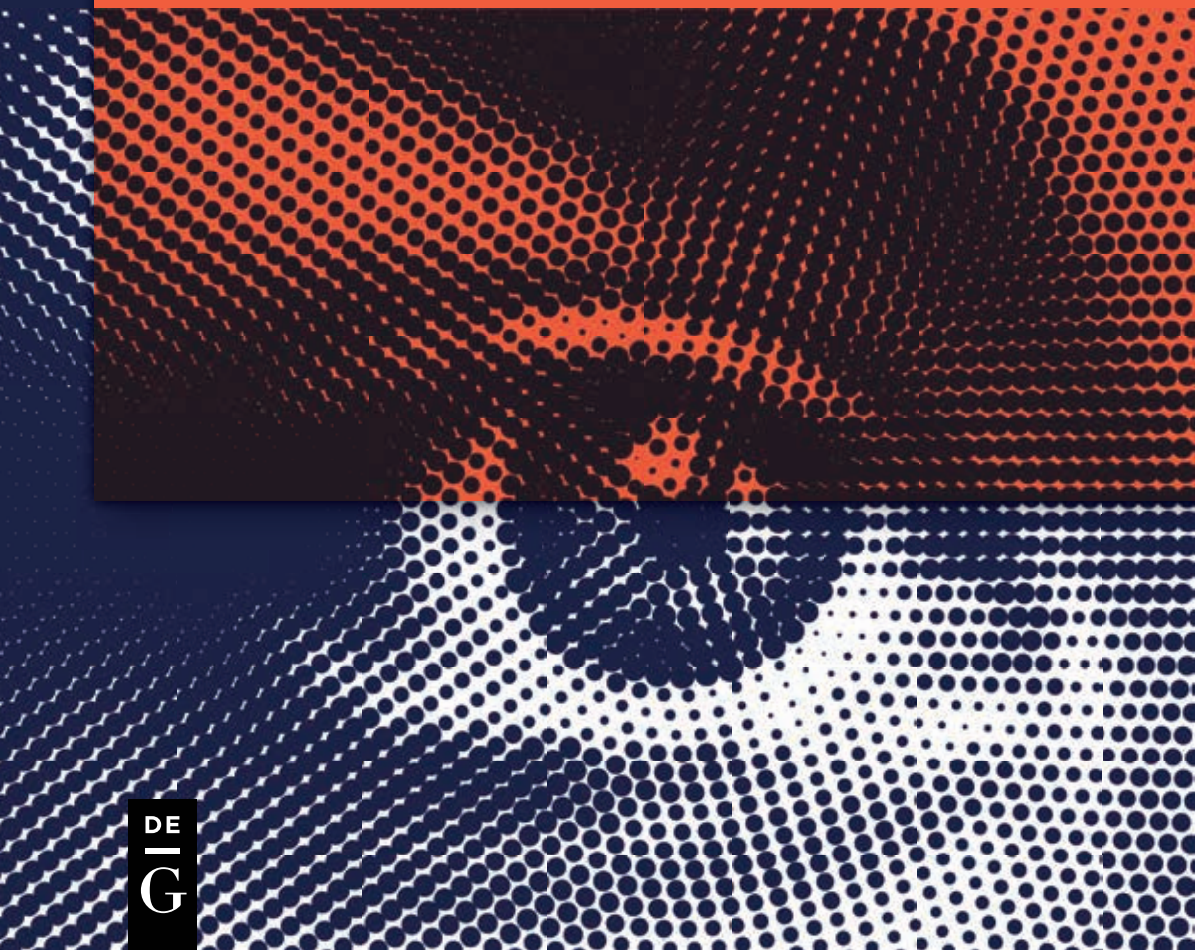
DE GRUYTER

Robert Hanna

DIGITAL TECHNOLOGY FOR HUMANS

THE MYTH OF AI, HUMAN DIGNITY, AND NEO-LUDDISM



Robert Hanna
Digital Technology for Humans

Robert Hanna

Digital Technology for Humans

The Myth of AI, Human Dignity, and Neo-Luddism

DE GRUYTER

ISBN 978-3-11-125979-6
e-ISBN (PDF) 978-3-11-126052-5
e-ISBN (EPUB) 978-3-11-126119-5

Library of Congress Control Number: 2025935127

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

© 2025 Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin
Cover image: cundra / iStock / Getty Images Plus
Printing and binding: CPI books GmbH, Leck

www.degruyter.com
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com

Acknowledgments

Earlier versions of parts of the material used in this book were presented as a plenary lecture at the "Kant and the Ethics of Enlightenment: Historical Roots and Contemporary Relevance" Conference, Immanuel Kant Baltic Federal University, Kalinigrad, RU, in April 2019, and also at the Online Summer School on *Kantian Political Philosophy Today: Global Challenges*, European Consortium for Political Research, Keele University, UK and Jagiellonian University, PL, in October 2021. I'm grateful to the organizers of both events for inviting me, and to the members of both audiences for their good comments and questions.

Parts of section 2.2 have been adapted from this previously published article:

> "Essentially Embodied Kantian Selves and The Fantasy of Transhuman Selves," *Studies in Transcendental Philosophy* 3 (2023). Available online at URL = <https://ras.jes.su/transcendental/s271326680021060-6-1-en>, last accessed 8 January 2025.

Parts of section 2.5 have been adapted from this article:

> "Caveat Lector: From Wittgenstein to The Philosophy of Reading." *Journal for the Philosophy of Language, Mind and the Arts* 5 (2024): 249 – 274. Available online: DOI 10.30687/Jolma/2723 – 9640/2024/03/013.

Parts of section 2.9 have been adapted from this previously published article:

> "The Myth of AI, Existential Threat, Why the Myth Persists, and What Is to Be Done About It." *Borderless Philosophy* 7 (2024): 35 – 61. Available online at URL = <https://www.cckp.space/single-post/bp-7-2024-robert-hanna-the-myth-of-ai-existential-threat-why-the-myth-persists-and-what-is-to>, last accessed 8 January 2025.

And chapter 5 is based on this previously published article:

> "Philosophical Foundations for Digital Ethics and AI Ethics: A Dignitarian Approach." Co-authored with Emre Kazim. *AI and Ethics*. 26 February 2021. Available online at URL = <https://link.springer.com/article/10.1007/s43681-021-00040-9>, last accessed 8 January 2025.

https://doi.org/10.1515/9783111260525-001

I'm also grateful to my editors at De Gruyter for their expert assistance at each stage in the publication process: Christoph Schirmer, Inga Lassen, and Ulla Schmidt.

I'm grateful to the three journal editors and to the *AI and Ethics* article's co-author, Emre Kazim, for their permissions to use this material here.

Finally, I'm very grateful to Scott Heftler for helping me have a "conversation" with ChatGPT and for thought-provoking conversations on and around the topics of section 2.3; to Mark Pittenger for thought-provoking conversations on and around the topics of section 2.8; and also to Otto Paans for thought-provoking correspondence on and around the topics of chapters 6 and 7.

Contents

A Note on References to Kant's Works —— IX

Further Abbreviations —— XI

1 Introduction: Digital Technology Is Our Tool, Not Our Master —— 1

2 Beyond Deep Thought: A New Critique of AI —— 13
2.1 Introduction —— **13**
2.2 Turing, Strong AI, and the Fantasy of Transhumanist Spiritualism —— **14**
2.3 How and Why ChatGPT Failed The Turing Test —— **20**
2.4 It's All Done with Mirrors: A New Argument That Strong AI Is Impossible —— **28**
2.5 Are There Some Legible Texts That Even the World's Most Sophisticated Robot Can't Read? —— **30**
2.6 How and Why to Perform Uncomputable Functions —— **44**
2.7 Babbage-In, Babbage-Out: On Babbage's Principle —— **48**
2.8 Creative Rage Against the Computing Machine: Necessary and Sufficient Conditions for Authentic Human Creativity —— **51**
2.9 The Myth of AI, Existential Threat, and Why the Myth Persists —— **63**
2.10 Organoid Intelligence? Just Say *No* —— **82**

3 What's So Special About Human Dignity? Its Metaphysics —— 87
3.1 Introduction —— **87**
3.2 How to Refute the Dignity-Skeptic —— **90**
3.3 What Is Human Dignity? —— **90**
3.4 The Metaphysical Ground of Human Dignity: Human Real Personhood —— **96**
3.5 Frederick Douglass, Kant, and Human Dignity —— **104**

4 Human Dignity and The Highest Good: Nonideal Dignitarian Moral Theory —— 108
4.1 Introduction —— **108**
4.2 The Thin Logic and the Thick Semantics of Moral Principles in Broadly Kantian Nonideal Dignitarian Moral Theory —— **111**
4.3 How to Solve the Universalizability and Rigorism Problems —— **119**
4.4 How to Solve the Problem of Moral Dilemmas —— **130**

4.5 Policy of Truth: The Murderer-at-the-Door Revisited — **141**
4.6 One Last Thing, by Way of Concluding This Chapter — **146**

5 Moral Principles of Dignitarian Digital Ethics — 149
5.1 Introduction — **149**
5.2 Four Approaches to Digital Ethics — **149**
5.3 Our Moral Obligations with Respect to Humanity — **155**
5.4 Moral Obligations with Respect to the Design, Creation, and Use of Digital Technology — **157**
5.5 What Privacy Is, and Why Invasions of Digital Privacy Are Morally Impermissible, Whereas Consensual Entrances into Digital Privacy Are Either Morally Permissible or Even Obligatory — **158**
5.6 Dignitarian Morality *Versus* Legality, and DDE — **160**

6 How Digital Technology Can Systematically Shape Our Lives in Bad, False, and Wrong Ways, and How Radically Enlightened Dignitarians Can and Should Push Back Against It — 163
6.1 Introduction — **163**
6.2 The Life-Shaping Thesis — **164**
6.3 John von Neumann and Richard Feynman — **166**
6.4 Nuclear Weapons Design and Virtual Reality — **167**
6.5 US Air Force Killer-Drone Operators — **168**
6.6 The New Wheel of Ixion: Work, Sleep, and the Mechanization of Everyday Life — **169**
6.7 From Classical Enlightenment to Cybernetic Radical Enlightenment — **174**
6.8 Cybernetic Radical Enlightenment, Stage I: Sociopolitical Critique, Logging Off, and Subverting-Dismantling-&-Replacing — **178**
6.9 Cybernetic Radical Enlightenment, Stage II: Civil Disobedience — **179**
6.10 Cybernetic Radical Enlightenment, Stage III: Developing Our Own Innate Capacities — **183**

7 Conclusion: Dignitarian Neo-Luddism with Respect to Digital Technology — 186

References — 190

Index — 202

A Note on References to Kant's Works

For convenience, I cite Kant's works in parentheses. The citations include both an abbreviation of the English title and the corresponding volume and page numbers in the standard "Akademie" edition of Kant's works: *Kants gesammelte Schriften*, edited by the Königlich Preussischen (now Deutsche) Akademie der Wissenschaften (Berlin: G. Reimer [now De Gruyter], 1902–). I generally follow the standard English translations of Kant's works, but have occasionally modified them where appropriate. For references to the first *Critique*, I follow the common practice of giving page numbers from the A (1781) and B (1787) German editions only. Because the Akademie edition contains only the B edition of the first *Critique*, I have also consulted the following German composite edition: *Kritik der reinen Vernunft*, ed. W. Weischedel, Immanuel Kant Werkausgabe III (Frankfurt: Suhrkamp, 1968). Here is a list of the relevant abbreviations and English translations:

CPR *Critique of Pure Reason.* Trans. P. Guyer and A. Wood. Cambridge: Cambridge Univ. Press, 1997.

CPrR *Critique of Practical Reason.* Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy.* Cambridge: Cambridge Univ. Press, 1996, pp. 139 – 272.

DSS "Dreams of a Spirit Seer Elucidated by Dreams of Metaphysics." In *Immanuel Kant: Theoretical Philosophy, 1755 – 1770.* Cambridge: Cambridge Univ. Press, 1992, pp. 301 – 359.

GMM *Groundwork of the Metaphysics of Morals.* Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy,* pp. 43 – 108.

JL "The Jäsche Logic." Trans. J.M. Young. In *Immanuel Kant: Lectures on Logic.* Cambridge: Cambridge Univ. Press, 1992, pp. 519 – 640.

LE *Immanuel Kant: Lectures on Ethics.* Trans. P. Heath. Cambridge: Cambridge Univ. Press, 1997.

MM *Metaphysics of Morals.* Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy,* pp. 365 – 603.

OT "What Does It Mean to Orient Oneself in Thinking?" Trans. A. Wood. In *Immanuel Kant: Religion and Rational Theology.* Cambridge: Cambridge Univ. Press, 1996, pp. 7 – 18.

Rel *Religion Within the Boundaries of Mere Reason.* Trans. A. Wood and G. di Giovanni. In *Immanuel Kant: Religion and Rational Theology,* pp. 57 – 215.

RTL "On a Supposed Right to Lie from Philanthropy." Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy,* pp. 611 – 615.

WiE "An Answer to the Question: What Is Enlightenment?" Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy,* pp. 17 – 22.

https://doi.org/10.1515/9783111260525-002

Further Abbreviations

AGI	Artificial General Intelligence
AHC	Authentic Human Creativity
AI	Artificial Intelligence
ASL	American Sign Language
BIBO	Babbage-In, Babbage-Out Principle
BIV	Brain-in-a-Vat
DCI	Dignitarian Cybernetic Instrumentalism
DDDR	Daily Dignitarian Digital Refusards
DDE	Dignitarian Digital Ethics
EEST	Essentially Embodied Selves Thesis
EL	Enlightenment Lite
FA	Formula of Autonomy
FHE	Formula of Humanity as End-in-Itself
FRE	Formula of the Realm of Ends
FUL	Formula of Universal Law
FULN	Formula of the Universal Law of Nature
GIGO	Garbage-In, Garbage-Out Principle
GPT	Generative Pre-Trained Transformer
ICR	Intelligent Character Recognition
LaMDA	Language Model for Dialog Applications
LLM	Large Language Model
MLK	Martin Luther King Jr
MP	Moral principle
NLP	Natural Language Processing
NSA	National Security Agency
OCR	Optical Character Recognition
OI	Organoid Intelligence
RE	Radical Enlightenment
RLHF	Reinforcement Learning from Human Feedback
TTS	Theory of Thought-Shapers
USAF	United States Air Force
ZFC	Zermelo-Fraenkel set theory plus the Axiom of Choice

https://doi.org/10.1515/9783111260525-003

1 Introduction: Digital Technology Is Our Tool, Not Our Master

> These are some of the reasons why robots, which may finally be here, aren't—and really shouldn't be—a harbinger of something negative. In theory, more technology will mean more efficiency for America's employers and, perhaps, in a post-pandemic world, a safer one. The early signals, however, are that we should not fear new robot overlords. Instead, the sharpest, most strategic companies are making a simultaneous investment in people who can power progress. (Horn and Jackson, 2021)

> In the realm of ends everything has either a *price* or a *dignity* [*Würde*]. What has a price can be replaced by something else as its *equivalent*; what on the other hand is raised above all price and therefore admits of no equivalent has dignity. What is related to general human inclinations and needs has a *market price*; that which, even without presupposing a need, conforms with a certain taste, that is, with a delight in the mere purposeless play of our mental powers, has an *affective price* [*Affectionspreis*]; but that which constitutes the condition under which alone something can be an end in itself has not merely a relative worth, that is, a price, but an inner worth, that is, *dignity*. Now, morality is the condition under which alone a rational being can be an end in itself, since only through this is it possible to be a lawgiving member in the realm of ends. Hence morality, and humanity insofar as it is capable of morality, is that which alone has dignity. (*GMM* 4: 434–435)

Is digital technology—including computers, algorithms, digital data or information (Adriaans, 2020), artificial intelligence/AI, and robotics—our *tool* or our *master?* I'll call the first option *cybernetic*[1] *instrumentalism,* and the second option *cybernetic oppression*, up to and including *cybernetic totalitarianism*—as, for example, brilliantly imagined in Jean-Luc Godard's 1965 dystopian science fiction classic, *Alphaville* (Godard, 1965).

It's essential to understand from the outset, however, that the problem posed by *Alphaville* is *not* a fundamental problem in the metaphysics of mind. Clearly and distinctly, the fictional mega-computer Alpha-60 is *neither* alive, *nor* conscious (i. e., subjectively experiencing), *nor* self-conscious (i. e., conscious of its own consciousness, second-order conscious), *nor* caring (i. e., desiring, emoting, or feeling), *nor* sensibly cognitive (i. e., sense-perceiving, remembering, or imagining), *nor* intellectually cognitive (i. e., conceptualizing, believing, judging, or inferring), *nor* volitional (i. e., deciding, choosing, or willing), *nor* intentionally object-directed or act-directed more generally, *nor* does it possess free agency (i. e., free will and practi-

1 The *Oxford Encyclopedic English Dictionary* defines "cybernetics" as "the scientific study of communications and automatic control systems in both machines and living things" (Hawkins and Allen, 1991: p. 358). Correspondingly, in this book I'm using the adjective "cybernetic" in this general sense, without a commitment to any particular *theory* of cybernetics.

https://doi.org/10.1515/9783111260525-004

cal agency): hence it's *not* rational (see, e.g., Hanna, 2015a, 2018a, 2018b, 2018c, 2020a). Therefore the fictional Alpha-60 is *artificial*, in the sense that it's a (Turing-)machine; but at the same time, it's *not* intelligent, precisely because, like all (Turing-)machines, by its very nature, it simply *cannot* "think" in the sense that all ordinary rational human minded animals can and do think (see sections 2.2 to 2.10 below). So, "artificial intelligence" *is an oxymoron.*

Moreover, despite the fact that the very idea of Artificial Intelligence, aka AI, has been around for more than 70 years (see, e.g., Turing, 1950), the recent and contemporary furor about AI has been primarily generated by the much-publicized *invasion of the chatbots:* Google's LaMDA in 2021 and Bard in early 2023, Microsoft's Sydney in 2021, and especially OpenAI's ChatGPT in late 2022 (see section 2.3 below; see also Chomsky, Roberts, and Watamull, 2023). But just like the fictional Alpha-60, although all the real-world chatbots are artificial, *none* of them are actually intelligent in the sense that rational human minded animals like us are intelligent. No, the problem posed by *Alphaville* is the fact that what I'll call *the military-industrial-digital complex*[2] (see, e.g., Harper's, 2024) is terrifyingly authoritarian, coercive, and brutal. This digitally-driven military-industrial complex, represented in the movie by the character Professor von Braun, who controls not only Alpha-60, but also the totalitarian State of Alphaville—in effect, Paris circa 1965, in the wake of the bloody Algerian War for independence from France, from 1954 to 1962—controls the world by creating and managing automated weapons systems and nuclear warfare, by universal surveillance, and above all by manipulating and mechanizing people's lives everywhere through all-pervasive and ubiquitous digital technologies. Now jumping forward to the third decade of the 21st century, in which the threat of a globalized version of the totalitarian State of Alphaville has become all-too-real, let's call this *the problem of digital technology.*

2 This riffs on a famous phrase in US President Dwight D. Eisenhower's "Farewell Address" in 1961: "[The] conjunction of an immense military establishment and a large arms industry is new in the American experience. The total influence—economic, political, even spiritual—is felt in every city, every statehouse, every office of the federal government. We recognize the imperative need for this development. Yet we must not fail to comprehend its grave implications. Our toil, resources and livelihood are all involved; so is the very structure of our society. In the councils of government, *we must guard against the acquisition of unwarranted influence, whether sought or unsought, by the military-industrial complex.* The potential for the disastrous rise of misplaced power exists, and will persist. We must never let the weight of this combination endanger our liberties or democratic processes. We should take nothing for granted. Only an alert and knowledgeable citizenry can compel the proper meshing of the huge industrial and military machinery of defense with our peaceful methods and goals so that security and liberty may prosper together." (See, e.g., Wikipedia, 2025b, italics added)

So, although the problem of digital technology is *not* ultimately a fundamental problem in the metaphysics of mind, it *is* a fundamental moral, sociopolitical, and existential-spiritual problem, even despite the soothing, sophistical reassurances of contemporary consultant experts in digital ethics (see, e.g., Muldoon, 2021) who work either directly for the technocratic global business corporations that reap immense profits from digital technology—the corporation OpenAI was instantly valued at *29 billion USD* on the day that it released ChatGPT in November 2022 (Wikipedia, 2025a)—or for university-based digital ethics "centers," "institutes," or "programs" that are heavily funded by those very same corporations.

For example, here's how M.B. Horn and C.J. Jackson purport to dispel what they call the "apparent paradox" about digital technology and the lives of workers:

> Scientist and fiction writer Isaac Asimov once wrote, "You can't differentiate between a robot and the very best of humans." The fear of automation, if not robot overlords, has been a fixation of business analyses and science fiction for generations. So with U.S. employers embarking on a clearer path toward machines, it might be easy to read these developments as a dire sign for American workers. In particular, some victims of robots would seem to be investments in areas like upskilling and education. The presumption might be that building a better bot gives you more bang for your buck.
>
> Except the opposite appears to be happening. Automation is here, but so too is a deeper appreciation for and investment in things like upskilling, learning and development, and education for all workers. (Horn and Jackson, 2021)

So, according to Horn and Jackson, any metaphysical, moral, sociopolitical, or existential worries we might have about digital technology are to be bought off by technocratic global business corporations investing big or not-so-big bucks in (re-)educating an elite sub-class of workers at those self-same technocratic global business corporations. Nice, very nice.

My view is radically different from and sharply opposed to this way of thinking about digital technology. More specifically, the core thesis of what I call Dignitarian Digital Ethics, or DDE, is that digital technology, no matter how powerful, sophisticated, or profitable, is nothing more and nothing less than a tool created by humankind for the betterment of humankind, whose use therefore not only can but also should be strictly constrained by general and specific moral principles flowing from the concept and fact of human dignity. I call this core thesis Dignitarian Cybernetic Instrumentalism, or DCI. Now all tools significantly shape the minds and lives of the users of those tools. Moreover, all high-powered tools can be used either well or wrongly—to nudge us, control us, torture us, or kill us; and they can also accidentally spin out of control, and cause serious damage, catastrophe, or even apocalypse. Therefore, DCI is a maximally morally robust version of cybernetic instrumentalism that's explicitly and fully opposed to cybernetic op-

pression of any and all kinds, intentional or accidental, up to and including cybernetic totalitarianism. The approach to digital technology that's proposed by DDE and DCI is, in effect, a dignitarian neo-Luddism with respect to digital technology.[3] Or in other and fewer words: digital technology only within the limits of human dignity. Correspondingly, dignitarian neo-Luddism with respect to digital technology is a maximally morally robust version of what's called *anthropocentrism* in ethics in general and digital ethics in particular (see, e.g., Torrance, 2011; see also section 5.2 below).

In order to clarify and explicate these claims, here are some definitions.

By *consciousness*, I mean the capacity for immanently reflexively aware, egocentrically-centered—aka subjective—essentially embodied experience that's innately possessed by rational human minded animals and also by other non-rational minded animals, whether human or non-human.

Rational human minded animals and other conscious animals are all capable of *intentionality*—aka the "aboutness" or "directedness" of conscious mental acts, states, or processes—which is the conscious mental representation of any object in the world, or of any other aspect of that world, of any act to be carried out by that minded animal, or of themselves (i.e., self-consciousness). Intentionality is a basic fact about us and other minded animals, that's exemplified by any minded animals of any species that are capable of any or all of these: consciousness, self-consciousness, caring, sensible cognition, intellectual cognition, volition, and free agency. In particular, if you're actually reading these words now, then you're capable of basic intentionality.

By saying that a fact or phenomenon is *basic*, I mean (i) that it's an irreducible and essential feature of at least some, many, or even all entities that are manifestly real, (ii) that it's fully capable of grounding other non-basic facts or phenomena, and (iii) that it's neither unanalyzable, nor non-relational, nor inexplicable.

Correspondingly, by contrast with basic intentionality, some representations are not, as items in the manifestly real world, in-and-of-themselves, conscious mental representations with basic intentionality, although they're actually used, or at least in principle can be used, by conscious animals as representations

3 See also Glendinning, 1990. I emphasize, and indeed re-emphasize, that in this book I'm not advocating neo-Luddism with respect to all technology, but only with respect to digital technology. Nevertheless, there are some overlaps between neo-Luddite approaches to environmental and political philosophy, and my own views (Hanna and Paans, 2022); and I'm also committed to a thoroughgoing rejection of the mechanistic worldview and to its revolutionary replacement by the neo-organicist worldview (Hanna, 2024b). See also chapter 7 below for a fully explicit definition of "dignitarian neo-Luddism with respect to digital technology."

with non-basic intentionality.[4] Leading examples of representations with non-basic intentionality include human language, logic, mathematics, diagrams, pictures, and symbols more generally.

By *data or information*, I mean the communicative, semantic, logical, mathematical, or other kind of representational content of any kind of representation —i. e., what that representation denotes, connotes, expresses, indicates, means, presents, says, etc., via its basic or non-basic intentionality—as opposed to that representation's encoding or presentational format alone.

And by *digital data or information*, I mean data or information that's actually or potentially encoded or presentationally formatted as a binary sequence of 0s and 1s, and can be operated on by the primitive recursive functions of arithmetic.

Then, by *digital technology* I mean the following:

> Anything *X* counts as "digital technology" if and only if *X* is an abstract or real-world machine, or a sub-part of such a machine, that operates on and/or processes digital data or information according to the logico-mathematical principles of Turing-computation (Turing, 1936/1937; Boolos and Jeffrey, 1989), aimed at some end or purpose.

In turn, by Artificial Intelligence, aka AI—a term, as I mentioned above, that turns out to be an oxymoron—I mean a proper subset of digital technology that's designed specifically in order to mimic or model human consciousness, self-consciousness, caring, sensible cognition, intellectual cognition, volition, intentionality more generally, and free agency, up to the point at which it can also greatly exceed contextual or natural limitations on human performance, or even fool rational human animals into believing that they're encountering another mind (Turing, 1950). In turn, AI includes robotics, since robots are real-world implementations of some or another version of AI—even when robots are so-called "autonomous," in the attenuated sense that they're more or less detached from central control systems, i. e., they're mobile, and operating "in the wild."

---

4 The distinction I'm making here between basic intentionality and non-basic intentionality is similar in some ways to John Searle's distinction between intrinsic intentionality and derived intentionality. See, e. g., Searle, 1980a, 1983, 1992: esp. pp. 78 – 82. The crucial differences between my distinction and Searle's are (i) that my distinction presupposes the essential embodiment theory of the mind-body relation (Hanna and Maiese, 2009; Hanna, 2011a), whereas by sharp contrast Searle's distinction presupposes a biologically-driven and causal version of non-reductive materialism/physicalism, and (ii) although my notion of intentionality's basicness rules out any implication of its unanalyzability, non-relationality, and inexplicability from the get-go, Searle's notion of intentionality's intrinsicness doesn't rule this out.

Now an *algorithm*, is any well-defined, finite, data-or-information-processing routine or rule-governed sequence, according to the logico-mathematical principles of Turing-computation. Algorithms are implemented by digital technology that's specifically in the form of real-world artificial automata or artificial machines, aka "computers," and also by digital technology more generally, of many different kinds, built out of many different kinds of materials. The end or purpose of any algorithm—say, to solve a class of problems, or to perform a calculation, or to yield some other sort of result (aka "optimization")—is pre-established by humankind, i. e., by us.

Therefore, all forms of digital technology, including computers, algorithms, digital data or information, AI, and robotics, are nothing more and nothing less than tools for data-or information-processing, or for real-world implementations of these, according to the logico-mathematical principles of Turing-computation, created by humankind for specifically human ends and purposes. Again, that's the thesis of cybernetic instrumentalism.

By *ethics*, I mean the domain of humankind's basic individual and social commitments and its leading ideals and values. And by *morality*, I mean humankind's attempt to guide human choice and action by rationally formulating and following principles or rules that reflect humankind's basic individual and social commitments and its leading ideals and values; and morality is the *core* of ethics.

Therefore, by *digital ethics* I mean humankind's attempt to guide human choice and action in the design, creation, and use of digital technology, including computers, digital data or information, AI, and robotics, by rationally formulating and following moral principles or rules that reflect humankind's basic individual and social commitments and its leading ideals and values.

Sadly, popular moral thinking and sociopolitical thinking are rife with fallacies, and contemporary digital ethics is no exception. Here's the general form of a particularly pernicious form of argument that I'll call *The Fallacy of Inevitability:*

(i) X is a very large social institution (in terms of the number of people who are members of that institution), X is very profitable for an elite group of powerful people, X has been in existence for a very long time (let's say, anywhere from 50 years to hundreds or even thousands of years), and X is very widespread (whether locally, regionally, nationally or even globally),
(ii) therefore X is inevitable, and
(iii) because X is inevitable, therefore X ought to be accepted by us *even if* X itself and its consequences are very often bad, false, and wrong, and furthermore,
(iv) therefore the most we can do in order to respond to the badness, falsity, and wrongness of X is to impose legal audits, regulations, or restrictions on X, while still allowing X to exist essentially unchanged in its current form.

In turn, let's consider some examples of real-world values of X:

X = chattel slavery
X = people's owning, carrying, or using guns
X = the State's owning, carrying, or using guns and/or weapons more generally, especially including nuclear weapons
X = military and domestic "intelligence," i.e., spying, together with policing, together with coercive authoritarian government, aka "the surveillance State," aka "the security State"
X = ecologically-destructive, technocratic global corporate capitalism
X = any digital technology whose design, creation, and/or use fails to treat people with sufficient respect for their human dignity (see also, e.g., McDonald et al., 2021), up to and including cybernetic totalitarianism

Now there's not merely one fallacy, but in fact three sub-fallacies, contained under The Fallacy of Inevitability, that I'll call *the three modes* of The Fallacy.

First, the step from from (i) to (ii) is clearly a *non sequitur.* Just because a social institution is very large, very profitable for an elite group of powerful people, has been in existence for a very long time, and is very widespread, it does *not* follow that this social institution is inevitable since no social institution is necessitated either by the laws of logic or the laws of nature. More generally, because humankind freely creates and freely sustains all actual and possible social institutions, then humankind can also freely abolish, refuse, or at the very least radically devolve-and-replace (hence, transform into its moral opposite) any such social institution. That's mode 1 of The Fallacy.

Second, step (iii) clearly contains another *non sequitur.* Since social institutions are contingent facts about humankind, but no contingent fact automatically entails a moral obligation, then step (iii) is clearly an instance of the naturalistic fallacy, namely, arguing directly from the factual (the "is") to the morally obligatory (the "ought"). Indeed, and diametrically on the contrary, if any social institution itself and its consequences are bad, false, and wrong, then humanity ought freely to abolish it, refuse it, or at the very least radically transform it into its moral opposite. That's mode 2 of The Fallacy.

Third, step (iv) is yet another clear *non sequitur.* On the assumption that some social institution itself and its consequences are bad, false and wrong, then it simply does not follow that the most we can do in order to respond to the badness, falsity, and wrongness of X is to impose legal audits, regulations, or restrictions on X, while still allowing X to exist essentially unchanged in its current form. Diametrically on the contrary, humankind ought freely to abolish it, refuse it, or at the

very least radically transform it into its moral opposite. And that's mode 3 of The Fallacy.

Consider, for example, chattel slavery, a social institution that's been in existence since even before the emergence of the earliest States (Scott, 2017: p. 155), and during certain historical periods has been practiced virtually worldwide. Chattel slavery is self-evidently inherently bad, false, and wrong, and its consequences are bad, false, and wrong, precisely because it violates human dignity—and for the basics of a theory of human dignity, see chapters 3 and 4 below.[5] And this holds even if chattel slavery is a very large social institution (say, forcibly employing and/or non-forcibly servicing millions of people living in the USA), very profitable for an elite group of powerful people (say, tobacco plantation owners and cotton plantation owners, and their business affiliates, in the American South), has been in existence for a very long time (say, from 1776 to 1865, and also even earlier during the pre-Revolutionary period), and is very widespread (say, spread all across the American South during those periods). Chattel slavery wasn't and isn't inevitable, and it wasn't and isn't even morally impermissible, much less morally obligatory, precisely because it violates human dignity, even when it was an actual social fact. Hence it would have been a moral absurdity and a moral scandal, merely to impose legal audits, regulations, and restrictions on chattel slavery, since the absurdity and moral scandal of the very idea of "common sense slavery control" are self-evident. Diametrically on the contrary, and also self-evidently, it was, is, and forever will be humankind's moral obligation to abolish chattel slavery, refuse it, or at the very least radically transform it into its moral opposite. And indeed, that's what actually happened to chattel slavery in the USA by the end of The Civil War—although, catastrophically and tragically, as everyone knows, or at least should know, the USA has continued to suffer for another 160 years from persistent and pervasive racist violations of human dignity and other malign consequences of structural or systematic racism, aka "white rage" (Anderson, 2017), from Reconstruction and the Jim Crow period, through the Civil Rights era and its troubled aftermath in the 1970s to the end of the 20th century, and now into the Black Lives Matter era and through the first two decades of the 21st century, right up to 6 am this morning.

Finally, to come to the sticking point, here's an excerpt from a 2021 article, "The Algorithmic Auditing Trap," by another consultant expert in digital ethics, Mona Sloane:

5 And for a full elaboration of the theory, including grappling with various hard cases, see Hanna, 2018c.

We have a new A.I. race on our hands: the race to define and steer what it means to audit algorithms. Governing bodies know that they must come up with solutions to the disproportionate harm algorithms can inflict.

This technology has disproportionate impacts on racial minorities, the economically disadvantaged, womxn, and people with disabilities, with applications ranging from health care to welfare, hiring, and education. Here, algorithms often serve as statistical tools that analyze data about an individual to infer the likelihood of a future event—for example, the risk of becoming severely sick and needing medical care. This risk is quantified as a "risk score," a method that can also be found in the lending and insurance industries and serves as a basis for making a decision in the present, such as how resources are distributed and to whom.

Now, a potentially impactful approach is materializing on the horizon: algorithmic auditing, a fast-developing field in both research and application, birthing a new crop of startups offering different forms of "algorithmic audits" that promise to check algorithmic models for bias or legal compliance….

Recently, the issue of algorithmic auditing has become particularly relevant in the context of A.I. used in hiring. New York City policymakers are debating Int. 1894 – 2020, a proposed bill that would regulate the sale of automated employment decision-making tools. This bill calls for regular "bias audits" of automated hiring and employment tools.

These tools—résumé parsers, tools that purport to predict personality based on social media profiles or text written by the candidate, or computer vision technologies that analyze a candidate's "micro-expressions"—help companies maximize employee performance to gain a competitive advantage by helping them find the "right" candidate for the "right" job in a fast, cost-effective manner.

This is big business. The U.S. staffing and recruiting market, which includes firms that assist in recruiting new internal staff and those that directly provide temporary staff to fill specific functions (temporary or agency staffing), was worth $151.8 billion in 2019. In 2016, a company's average cost per hire was $4,129, according to the Society for Human Resource Management.

Automated hiring and employment tools will play a fundamental role in rebuilding local economies after the Covid-19 pandemic. For example, since March 2020, New Yorkers were more likely than the national average to live in a household affected by loss of income. The economic impact of the pandemic also materializes along racial lines: In June 2020, only 13.9 % of white New Yorkers were unemployed, compared to 23.7 % of Black residents, 22.7 % of Latinx residents, and 21.1 % of Asian residents.

Automated hiring tools will reshape how these communities regain access to employment and how local economies are rebuilt. Against that backdrop, it is important and laudable that policymakers are working to mandate algorithmic auditing. (Sloane, 2021)

This argument is a perfect example of The Fallacy of Inevitability in all three of its modes. No form of digital technology is inevitable,[6] even if it's very large in the social-institutional sense, very profitable for an elite group of powerful people, has been in existence a very long time (i. e., at least 50 years), and is very widespread. Moreover, humankind is under absolutely no moral obligation whatsoever to use any form of digital technology that violates sufficient respect for human dignity, just because it's an actual social fact. And above all, the assumption that the most we can do is to impose legal audits, regulations, and restrictions on digital technology is completely false. Although "common sense digital technology control" might not look morally scandalous and morally absurd at first glance, just as "common sense gun control" might not look morally scandalous and morally absurd at first glance, in fact they both are every bit as morally scandalous and morally absurd as "common sense slavery control."[7] Diametrically on the contrary, then, if any form of digital technology violates human dignity, then humankind not only can but should freely abolish, refuse, or at the very least radically transform it into its moral opposite. Period.—No matter how much this rejection disrupts the wonders of "innovation," the self-interests of digital technology billionaires and their global corporations, and/or the putatively public interests of the governments of neoliberal States.

To be fair, in "The Algorithmic Auditing Trap" Sloane does go on to say, immediately after the text I quoted earlier, that

> we are facing an underappreciated concern: To date, there is no clear definition of "algorithmic audit." Audits, which on their face sound rigorous, can end up as toothless reputation polishers, or even worse: *They can legitimize technologies that shouldn't even exist because they are based on dangerous pseudoscience.* (Sloane, 2021, italics added)

That emphasized sentence, at least prima facie, looks like the dignitarian abolish-it-refuse-it-or-radically-transform-it-into-its-moral-opposite principle I'm defending. But unfortunately, by the end of the article, she's fallen back into The Fallacy of Inevitability, by merely recommending "three steps": (i) "transparency about where and how algorithms and automated decision-making tools are deployed," (ii) "need[ing] to arrive at a clear definition of what 'independent audit' means

---

6 The false thesis of the inevitability of digital technology is sometimes called "technological determinism." But this term is also used in a baffling variety of other non-equivalent senses: see, e. g., Dafoe, 2015. Hence it's philosophically least confusing, and all-around better, simply to avoid using that term in this context altogether.

7 For an application of dignitarian moral reasoning to owning, carrying, or using guns, see Hanna, 2024a.

in the context of automated decision-making systems, algorithms, and A.I.," and (iii) "need[ing] to begin a conversation about how, realistically, algorithmic auditing can and must be operationalized in order to be effective." Morally speaking, that's pretty thin gruel.

Therefore, again: if digital technology is bad, false, and wrong because it violates sufficient respect for human dignity, then, just like chattel slavery, humankind not only *can* but *should* abolish it, refuse it, or at the very least radically transform it into its moral opposite. As I mentioned at the outset, the core thesis of Dignitarian Digital Ethics, aka DDE, is a radically robust version of ethical anthropocentrism, namely, Dignitarian Cybernetic Instrumentalism, aka DCI: digital technology, no matter how powerful, sophisticated, or profitable, is nothing more and nothing less than a tool created by humankind for the betterment of humankind, whose use therefore can and should be strictly constrained by general and specific moral principles flowing from the concept and fact of human dignity. Digital technology is our tool, not our master.

In turn, DCI is metaphysically grounded on what I call *the neo-organicist worldview* (see, e. g., Hanna and Maiese, 2009; Hanna and Paans, 2020; Torday, Miller Jr, and Hanna, 2020; Hanna, 2024b), which says (i) that the natural universe or cosmos is fundamentally uncomputable, negentropic, processual, purposive, self-organizing, and non-equilibrium thermodynamic, hence fundamentally non-mechanical (the organicity thesis), (ii) that all mechanical systems are explanatorily and ontologically derivative from organic systems, and not the other way around, as per classical or contemporary metaphysical materialism or physicalism and the Standard Models of cosmology and particle physics (the explanatory inversion thesis), and (iii) that there's a single, unbroken metaphysical continuity between The Big Bang Singularity, temporally asymmetric/unidirectional, uncomputable, negentropic, processual, purposive, self-organizing, non-equilibrium thermodynamic matter/energy flows, organismic life, minded animals generally, minded human animals specifically, their free agency, and their dignity (the mind-life continuity thesis). In diametric opposition to DCI and the neo-organicist worldview, any and all forms of cybernetic oppression, up to and including cybernetic totalitarianism, are grounded on what I call *the mechanistic worldview*, which says that everything in the world is fundamentally either a formal automaton or a natural automaton, operating strictly according to Turing-computable algorithms and/or time-reversible or time-symmetric deterministic or indeterministic laws of nature, especially the Conservation Laws (including the First Law of Thermodynamics) and the Second Law of Thermodynamics, which also imposes always-increasing entropy—i. e., the always-increasing unavailability of any system's thermal energy for conversion into causal (aka "mechanical") action or work—on all natural mecha-

nisms, until a total equilibrium state of the natural universe is finally reached (see also Hanna and Paans, 2020; Hanna, 2024b).

So, in view of the all-pervasiveness and ubiquity of digital technology in our contemporary and foreseeably future individual, social, economic, and political lives, it follows that DCI is of global metaphysical, moral, and existential significance—in all senses of "existential"—for humankind.

In the six chapters that follow this one, I'll present and defend the metaphysical foundations, moral principles, and the radically robust ethically anthropocentric sociopolitical and existential applications of DDE and DCI.

2 Beyond Deep Thought: A New Critique of AI

2.1 Introduction

In Douglas Adams's brilliantly satirical science fiction trilogy, *The Hitchhiker's Guide to the Galaxy* (Adams, 2002), the answer provided by the supercomputer Deep Thought, after 7.5 million years of computing, to the Ultimate Question of Life, the Universe, and Everything, is … "42." Mm, right. Sadly, on top of that, the Ultimate Question itself also remains unknown.

But speaking in all seriousness now, as a proponent of broadly Kantian dignitarian moral theory (see chapters 3–4 below), it seems self-evidently obvious to me that the Ultimate Question is just this:

> How do rational human minded animals, as essentially embodied, conscious, self-conscious, caring, sensibly cognitive intellectually cognitive, volitional, intentionally object-directed and act-directed free agents who inherently possess dignity, fit into the all-inclusive natural universe, aka the cosmos?

For if we can provide a fully adequate explanation of how we fit into the cosmos, then we'll thereby also have fully adequately explained Life, the Universe, and Everything. And that's because, in order to provide a fully adequate explanation of a proper part of a given whole, on the assumption that it's an integrated, systematic whole, you have to be able to provide a fully adequate explanation of the rest of that whole too. Making the assumption that the cosmos is an integrated, systematic whole, and according to the rationale I just spelled out, then, when we conjoin broadly Kantian dignitarian ethics with the neo-organicist worldview, we'll have a non-mathematical answer to the Ultimate Question; and therefore that conjunction is a theory of Life, the Universe, and Everything.

As we'll see in this chapter, not only supercomputers but also just plain old garden-variety non-supercomputer computers and/or robots, nowadays called "Artificial Intelligence" aka AI—or in the near future, even when powered by brain organoids grown from human brain cells cultivated in a dish or a vat, "Organoid Intelligence," aka OI—cannot answer the Ultimate Question, not even in principle, although according to the rationale I presented just above, we can, at least in principle. But this is only one of the ways we not only can, but in fact already do, go essentially beyond Deep Thought and all its kin; correspondingly, in this chapter I'll explore several of the principal ways, and thereby undertake what I call a *new* critique of AI.

https://doi.org/10.1515/9783111260525-005

2.2 Turing, Strong AI, and the Fantasy of Transhumanist Spiritualism

In 1933, under the heading of "[The] Nature of Spirit," Alan Turing wrote:

> [A]s regards the actual connection between spirit and body I consider that the body by reason of being a living body can "attract" and hold on to a "spirit" whilst the body is alive and awake and the two are firmly connected. When the body is asleep I cannot guess what happens but when the body dies the "mechanism" of the body, holding the spirit, is gone and the spirit finds a new body sooner or later perhaps immediately.
>
> As regards the question of why we have bodies at all; why we do not or cannot live free as spirits and communicate as such, we probably could do so but there would be nothing whatever to do. The body provides something for the spirit to look after and use. (As quoted in Popova, 2017)

Three years later, Turing published "On Computable Numbers, with an Application to the Entscheidungsproblem" (Turing, 1936/1937), the foundational text of the formal science of digital computing; and in 1950 he published "Computing Machinery and Intelligence," the foundational text of the AI program. It might not be immediately or intuitively obvious, but in fact there's an essential connection between what Turing wrote in 1933 on the one hand, and what he wrote in 1936/1937 and 1950 on the other, namely, what can be called *transhumanist spiritualism.*

What do I mean by that term? The recent and contemporary movement of *transhumanism* (aka "posthumanism") is a

> social and philosophical movement devoted to promoting the research and development of robust human-enhancement technologies. Such technologies would augment or increase human sensory reception, emotive ability, or cognitive capacity as well as radically improve human health and extend human life spans. Such modifications resulting from the addition of biological or physical technologies would be more or less permanent and integrated into the human body. (Britannica, 2025)

More specifically,

> [a]s humans, we are defined by, among other things, our desire to transcend our humanity. Mythology, religion, fiction and science offer different versions of this dream. Transhumanism—a social movement predicated on the belief that we can and should leave behind our biological condition by merging with technology—is a kind of feverish amalgamation of all four. Though it's oriented toward the future, and is fuelled by excitable speculation about the implications of the latest science and technology, its roots can be glimpsed in ancient stories like that of the Sumerian king Gilgamesh and his quest for immortality.... Transhumanism represents a desire to obliterate the boundary between human bodies and machines, and a confusion in the first place as to the distinction between the two. (O'Connell, 2018)

Even more specifically, transhumanism claims that the selves of creatures like us not only can exist independently of our bodies, as functional systems of representational content that are inherently able to be implemented or realized in digital technology and uploadable to servers, but also can survive accidental or natural human death in server-limbo, then be downloaded into technologically enhanced partially mechanical humanoid bodies or even into wholly artificially-created completely mechanical non-humanoid bodies, survive in these new implementations or realizations for an indefinitely long time, repeat that process, and possibly even become immortal. Now, it's true that the belief in and desire for digital-technological immortality isn't strictly required for transhumanism: as a recent survey showed, as many as 23.8% of contemporary transhumanists *don't* actually desire digital-technological immortality (Wikipedia, 2025c). But of course, that also implies that as many as 76.2% of contemporary transhumanists *do* actually desire (and therefore, presumably, also believe in) digital-mechanical immortality.

Leaving aside for a moment—although I will come back to it later in this section—transhumanism's inherently questionable metaphysical appeal to the mechanistic worldview, including the strong thesis of artificial intelligence, aka strong AI, and also its all-too-trendy, breathless, and high-gloss futurological appeal to the Promethean wonders of digital technology, transhumanism is in fact metaphysically equivalent to Swedenborgian spiritualism (see, e.g., Swedenborg Foundation, 2025), which Immanuel Kant so effectively criticizes and wittily derides in his 1766 book, *Dreams of a Spirit Seer Elucidated by Dreams of Metaphysics:*

> In Stockholm there dwells a certain [Swedenborg], a gentleman of comfortable means and independent position. For the last twenty years or more he has, as he tells us, devoted himself exclusively to cultivating the closest contact with spirits and with the souls of the dead, and, in exchange, to giving them information about this present world, to composing hefty volumes devoted to his discoveries, and periodically travelling to London in order to supervise their publication.... [Swedenborg] distinguishes between the outer and inner memory in humankind. A person has outer memory as someone belonging to the visible world, whereas a person has inner memory in connection with the spirit-world. [Swedenborg's] own superiority consists in the fact that, already in this life, he sees himself as a person who belongs to the community of spirits and that he is recognized as someone belonging to that community. It is also in this inner memory that everything, which has vanished from outer memory, is conserved, none of a person's representations ever getting lost. After death, the memory of everything which had ever entered his soul and which had so far remained concealed from him, goes to make up the complete book of his life. (*DSS* 2: 354, 362)

Now, by the term *essentially embodied rational human selves*, I mean necessarily and completely embodied, conscious, self-conscious, caring, sensibly cognitive, intellectually cognitive, volitional, intentionally object-directed and act-directed free-agential human animals, aka human persons, innately possessing dignity or

absolute value, and fully capable not only of free agency, but also of a priori knowledge of analytic and synthetic a priori truths alike, egocentrically centered in manifestly real orientable space and time. Or as Kant much more compactly puts it in *Dreams of a Spirit Seer:* "*My soul is wholly in my whole body, and wholly in each of its [organic] parts*" (*DSS* 2: 325, italics in the original, bracketted word taken from the original source quoted by Kant).

In my opinion, for which I've argued in detail and at length elsewhere, all rational but also "human, all too human" in Nietzsche's sense (by which I mean: animal, embodied, finite, and thoroughly normatively imperfect) creatures like us—for example, the readers of this book—are, synthetic a priori necessarily,[8] essentially embodied selves (Hanna and Maiese, 2009; Hanna, 2011a, 2011b, 2018b, 2023a). Let's call that The Essentially Embodied Selves Thesis, or for short, EEST, pronounced like "east." If EEST is true, then it's synthetic a priori impossible for the selves of creatures like us to exist independently of our own living organismic animal bodies or beyond the deaths of those bodies, as spiritual selves, whether temporarily or permanently, by any means whatsoever. Indeed, the very ideas of disembodied spiritual selves, their survival after death, and of human immortality, while minimally logically consistent, are in fact conceptually empty and incoherent, even over and above the synthetic a priori impossibility of such things, since the term "myself" indexically picks out an essentially embodied self, all of whose core features require grounding in a particular living organismic animal body. Moreover, and more importantly, if EEST is true, then, just like Swedenborgianism, transhumanist spiritualism is not only conceptually empty and incoherent, but also synthetic a priori impossible. And what's more, it's also existentially and morally reprehensible (Gare, 2016). In short, then, the belief in transhuman spiritual selves is nothing but an existentially and morally reprehensible metaphysical fantasy.

To be sure, Turing was an exceptionally brilliant and important mathematical logician and philosophical visionary, and also the decrypter of Nazi communications codes during World War II—thereby substantially contributing to shortening the war by 2–4 years and probably saving between 14 million and 21 million lives—who lived an all-too-short and tragic life (he committed suicide in 1954), intensely suffering under the moralistic coercive authoritarianism of the very State he so nobly served in World War II, England, because it also systematically persecuted homosexuals (see, e.g., Popova, 2017). Nevertheless, according to the conjunction of his own explicit formulations in 1933, 1936/1937, and 1950, Turing was also at

8 For a full presentation and detailed defense of the analytic-synthetic distinction, including the synthetic a priori, see Hanna, 2001, and 2015a: esp. ch. 4.

least implicitly committed to the fantasy of transhumanist spiritualism. Let me explain.

The classical Swedenborgian metaphysical fantasy about the self is that our "human, all too human" self can somehow exist in disembodied form, as a *Geist*/ghost-self, or spirit, that hovers epiphenomenally above the manifestly real world, immortally, yet also retains all the representational content that we ever personally or sub-personally cognized. The recent and contemporary crypto-Swedenborgian metaphysical fantasy of transhumanist spiritualism, post-Turing, is that this *Geist*/ghost-self or spirit is a functional system in digital format, as described by the formal science of digital computing and Turing machines (Turing, 1936/1937; Boolos and Jeffrey, 1989; Sipser, 2013), that can exist in abstract, disembodied form like a computer program, and be repeatedly uploaded or downloaded to, and implemented or realized on, on different kinds of digital technology, that also have other causal powers of various kinds, and thereby our transhuman spiritual self epiphenomenally hovers above the manifestly real natural and social world like a causally inert shadow, going from one uploading, downloading, implementation, or realization to another, escaping accidental or natural human death for an indefinitely long time and perhaps even forever in the sempiternal sense, i. e., immortally. Transhumanist spiritualism, in turn, presupposes the truth of the strong thesis of artificial intelligence, aka strong AI (for an explicit definition, see the next paragraph), which again in turn presupposes the truth of the materialist/physicalist doctrine of metaphysical computational-functionalism about the human mind, which says that the human mind and all its contents can be explanatorily and ontologically reduced to a Turing-computable system of input-output mappings that can be implemented or realized on different sorts of machines, all of which operate as digital computing machines, in addition to whatever other causal powers these machines might have. But if EEST is true, then both strong AI and metaphysical computational functionalism about the human mind are necessarily false.

More specifically, strong AI is the two-part thesis which says (i) that rational human intelligence can be explanatorily and ontologically reduced to Turing-computable algorithms and the operations of digital computers (aka the thesis of formal mechanism, as it's applied to rational human intelligence), and (ii) that it's technologically possible to build a digital computer that's an exact counterpart of rational human intelligence, such that this machine not only exactly reproduces (aka simulates) all the actual performances of rational human intelligence, but also outperforms it (aka the counterpart thesis) (see, e. g., Block, 1980: part 3; Kim, 2011: ch. 6). The strong AI thesis is not only immensely controversial, but also strongly apt to be seriously muddled, for at least three reasons.

First, the strong AI thesis is very often confused with the weak AI thesis, which says that performances of rational human intelligence can be "modeled" on digital computers, but (i) the weak AI thesis is itself ambiguous as between a non-trivial version and a trivial version, and (ii) even if both of the versions of the weak AI thesis were true, nevertheless strong AI could still be false.

Second, the strong AI thesis, as such, overlooks the fact that rational human intelligence is also conscious: hence if strong AI were true, then human consciousness would also have to be explanatorily and ontologically reducible to Turing-computable algorithms and the operations of digital computers, which is equivalent to metaphysical computational functionalism about human consciousness, which in turn puts an extra-heavy burden of proof on defenders of strong AI. Indeed, Ned Block, David Chalmers, Frank Jackson, and others have developed (admittedly, controversial) arguments which, if sound, show that phenomenal consciousness does not logically strongly supervene on functional physical facts (see, e. g., Chalmers, 1996: chs. 1–4). But even if, as I've argued elsewhere, the conception of phenomenal consciousness that's assumed by all those anti-functionalist arguments is false (Hanna and Maiese, 2009; Hanna, 2011a), nevertheless if EEST is true, then necessarily consciousness in creatures like us is essentially embodied, so metaphysical computational functionalism about human consciousness is false and indeed synthetic a priori impossible.

Third, the very idea of "being artificial" is ambiguous as between (i) being mechanical, as opposed to being organic, and (ii) being able to be built or constructed or synthesized, as opposed to not being able to be built or constructed or synthesized, for whatever reason, but (i) and (ii) are mutually logically independent of one another: something could be mechanical but not buildable, constructible, or synthesizable (for example, digital computations involving more digits or computations than there are particles or future moments of time in the cosmos), and conversely something could be buildable, constructible, or synthesizable but not mechanical (for example, certain exactly reproducible uncomputable, non-equilibrium thermodynamic biochemical processes, including organismic processes).

Over and above the controversies, the strong AI thesis is demonstrably false, for at least four good reasons, and also the weak AI thesis is either false (the non-trivial version) or boringly trivially true (the trivial version), as follows.

1. If EEST is true, then necessarily, intelligent rational human minds are alive, but systems that conform to the strong AI thesis are inherently mechanical and non-living, so the strong AI thesis is necessarily false.
2. If EEST is true, then necessarily, intelligent rational human minds are essentially embodied, but systems that conform to the strong AI thesis are possibly disembodied, so the strong AI thesis is necessarily false.

3. If EEST is true, then necessarily, rational human knowledge requires a non-accidental connection between judgment or belief and truth, and also a non-accidental connection between true belief and justification (Hanna, 2015a: esp. chs. 1 and 6 – 8), but systems that conform to the strong AI thesis only ever provide accidental content-connections, hence not only (i) is the strong AI thesis necessarily false, but also (ii) if the strong AI thesis were true, then our intelligent rational human minds would be nothing more than digital computers, and therefore we couldn't ever know the truth of the strong AI thesis, so the strong AI thesis is also epistemically self-undermining.
4. If EEST is true and also Kurt Gödel's incompleteness theorems (Gödel, 1931/1967), which say (4i) that all *Principia Mathematica*-style systems of mathematical logic based on the Peano axioms for arithmetic will contain undecidable, unprovable sentences, and (4ii) that no such system of mathematical logic can prove its own consistency, hence the truth of mathematical axioms has to be demonstrated outside those systems—for example, by acts of rational human mathematical intuition (see, e. g., Hanna, 2015a: chs. 6 – 8)—are true, then EEST together with Gödel's incompleteness theorems jointly entail that there will be uncomputable, undecidable, unprovable mathematical axioms that only intelligent rational human minded animals, i. e., essentially embodied selves, can know, so systems that conform to the strong AI thesis inherently fall short of the characteristic achievements of rational human intelligence, and therefore the strong AI thesis is false (see also Keller, 2023, and Landgrebe and Smith, 2022).

Moreover, if the weak AI thesis about "modeling" performances of rational human intelligence on digital computers, says that not all but only some actual performances of rational human intelligence are exactly reproducible (aka can be reincarnated) on digital computers (i. e., the non-trivial version), then since the strong AI thesis is not only false but impossible, then the non-trivial version of the weak AI thesis is false and impossible too. But if the weak AI thesis says merely that some behavioral or formal features of some actual performances of rational human intelligence are either operationally or isomorphically representable, i. e., able to be mimicked, on digital computers (the trivial version), then this is indeed true, but at best boringly trivially true, since the very same thesis is true of even the simplest counting or calculating procedures, using, for example, one's fingers, hockey pucks, or an abacus. In short, the very idea of transhumanist spiritual selves is nothing but a metaphysical fantasy—a digital *Geist*/ghost-self or spirit—of precisely the sort so effectively criticized and wittily derided by Kant in *Dreams of a Spirit Seer*. And as we've seen, if EEST is true, then the very idea of transhuman selves is conceptually empty or incoherent and synthetic a priori impossible.

Now, the only remaining philosophical question is, why are post-Turing transhumanists so attracted to, and indeed intellectually and emotionally addicted to, the crypto-Swedenborgian spiritualist fantasy that their own "human, all too human" selves are digital *Geist*/ghost-selves or spirits? By way of answering that question, I strongly agree with Arran Gare's critical diagnosis of the postmodernist/posthumanist "quest for disembodiment":

> The illusory quest for disembodiment … appears to be common among "macroparasites" who live off the work, products and lives of others. Historically, this illusion of disembodiment appears to have legitimated exploitation of others, but in doing so has led the ruling classes of civilizations to destroy the real conditions of their own existence. The … postmodernist forms of this illusion … are transmogrifications of the illusion of disembodiment on which medieval civilization was based. (Gare, 2016: p. 27)

So, assuming that Gare's critical diagnosis is correct, then not only are the transhumanist crypto-Swedenborgian fantasists of transhuman selves—digital *Geist*/ghost-selves or spirits—self-deniers of their own living organismic human animal embodiment (i. e., they're committed to an inherently inauthentic way of life by virtue of existential self-stultification), they're also explicitly or implicitly intellectually and emotionally addicted to a rationally unjustified and immoral coercive authoritarian sociopolitical theory (i. e., they're committed to a social-institutional system that violates the universal obligation to treat all rational human animals with sufficient respect for their human dignity, and is thereby inherently oppressive). The very idea of transhuman spiritual selves is therefore not only a crypto-Swedenborgian metaphysical fantasy, it's also an existentially and morally reprehensible fantasy. Now, it's an obvious and tragic truth that people can become addicted to things that are not only bad for them personally, but also objectively bad. And so it is with post-Turing transhumanist spiritualism.

2.3 How and Why ChatGPT Failed The Turing Test

As everyone who's interested in higher education knows, in late 2022 and early 2023 there was a brouhaha in contemporary higher education in particular and contemporary society more generally about ChatGPT (see, e. g., Boethius, 2022). But here I'm concerned with ChatGPT in relation to fundamental issues in the philosophy of mind, not with the brouhaha.

The OpenAI blog says this:

> ChatGPT: Optimizing Language Models for Dialogue. We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. (OpenAI, 2025a)

And Wikipedia says this:

> ChatGPT (Generative Pre-trained Transformer) is a chatbot launched by OpenAI in November 2022. It is built on top of OpenAI's GPT-3 family of large language models, and is fine-tuned (an approach to transfer learning) with both supervised and reinforcement learning techniques.
>
> ChatGPT was launched as a prototype on November 30, 2022, and quickly garnered attention for its detailed responses and articulate answers across many domains of knowledge. Its uneven factual accuracy was identified as a significant drawback. Following the release of ChatGPT, OpenAI was reportedly valued at $29 billion. (Wikipedia, 2025c)

Of course, the "AI" in "OpenAI" stands for "Artificial Intelligence," which should immediately remind us of Turing's famous essay, "Computing Machinery and Intelligence," which begins as follows:

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the "imitation game." (Turing, 1950: p. 433)

In other words, Turing is using "intelligence" as a synonym for "the capacity to think"; and what he calls "the imitation game" is the same as The Turing Test. Correspondingly, the Stanford Encyclopedia of Philosophy article on "The Turing Test" glosses Turing's argument as follows:

> The phrase "The Turing Test" is most properly used to refer to a proposal made by Turing (1950) as a way of dealing with the question whether machines can think. According to Turing, the question whether machines can think is itself "too meaningless" to deserve discussion (442). However, if we consider the more precise—and somehow related—question whether a digital computer can do well in a certain kind of game that Turing describes ("The Imitation Game"), then—at least in Turing's eyes—we do have a question that admits of precise discus-

> sion. Moreover, as we shall see, Turing himself thought that it would not be too long before we did have digital computers that could "do well" in the Imitation Game. The phrase "The Turing Test" is sometimes used more generally to refer to some kinds of behavioural tests for the presence of mind, or thought, or intelligence in putatively minded entities. (Oppy and Dowe, 2021)

Reduced to its essentials, The Turing Test deploys written natural language, in the form of questions and answers exchanged between two writers, one of whom is a rational human experimenter and the other a test subject, as a criterion for the presence or absence of what I'll call "rational human mindedness" in the test subject. By "rational human mindedness," I mean the possession of a unified set of seven innate capacities for consciousness, self-consciousness, caring, sensible cognition, intellectual cognition, volition, object-directed and act-directed intentionality more generally, and free agency, hence for rationality, including natural language competence (for more elaboration, see section 2.7 below). The term "rational human mindedness," as I'm using it, means the same as "human real personhood" (see, e. g., Hanna, 2018b: chs. 6 – 7). By "human" in this context, I mean essentially the same what Nietzsche meant by "human, all too human," namely, "finite, fallible, and thoroughly normatively imperfect." "Rational human mindedness," as I am understanding it, applies to beings that aren't necessarily *biologically* human, since rational human minded beings in alien or other earth-bound species are wholly meaningfully conceivable, as fantasy writing and science fiction clearly show. So, what's being tested-for in The Turing Test, is an ability to use written natural language in such a way as to convince an interlocutor who already possesses rational human mindedness, that the other interlocutor also possesses all of the capacities that are jointly constitutive of rational human mindedness, namely, consciousness, self-consciousness, caring, sensible cognition, intellectual cognition, volition, object-directed and act-directed intentionality more generally, and free agency.

Now, Turing himself apparently holds that something's passing The Turing Test in any given case is a sufficient condition for that thing's having rational human mindedness in that case. But that's obviously a mistake. Just because a machine can be created which writes in such a way that it convinces someone that it has rational human mindedness, it doesn't follow that the machine actually has rational human mindedness. Similarly, just because a machine can be created that convinces someone that it's God, or Satan, or The Wizard of Oz, it doesn't follow that the machine actually is God, or Satan, or The Wizard of Oz. Most people are naturally fairly gullible and trusting (I'm not saying that this is necessarily a bad thing, in fact I think it's often and even on the whole a very good thing—just that it's a fact), especially romantically-minded people, and therefore it's fairly

easy to deceive them, and convince them to believe things that are actually false or bad, as all bullshitters, dictators, would-be dictators, other demagogues, con men, other grifters, advertising-copy writers, influencers, and other sophists know full well; and this of course also applies directly to The Turing Test. Consider, for example, the Google LaMDA controversy:

> In June 2022 the Google LaMDA (Language Model for Dialog Applications) chatbot received widespread coverage regarding claims about it having achieved sentience. Initially in an article in The Economist Google Research Fellow Blaise Agüera y Arcas said the chatbot had demonstrated a degree of understanding of social relationships. Several days later, Google engineer Blake Lemoine claimed in an interview with the Washington Post that LaMDA had achieved sentience. Lemoine had been placed on leave by Google for internal assertions to this effect. Agüera y Arcas (a Google Vice President) and Jen Gennai (head of Responsible Innovation) had investigated the claims but dismissed them. Lemoine's assertion was roundly rejected by other experts in the field, pointing out that a language model appearing to mimic human conversation does not indicate that any intelligence is present behind it. (Wikipedia, 2025d)

Clearly, according to Blake Lemoine, LaMDA passed The Turing Test, yet Lemoine was simply deceived by LaMDA.

Nevertheless, even though passing The Turing Test in any given case isn't a sufficient condition for having rational human mindedness, clearly, failing The Turing Test in all cases is a sufficient condition for not having rational human mindedness. Given most people's natural gullibility and trustingness, if something doesn't use written natural language in a way that's sufficient to convince anyone who already has rational human mindedness and is also probably fairly gullible and trusting, that the other interlocutor has rational human mindedness, then obviously that thing doesn't have rational human mindedness, and therefore, in Turing's terminology, that thing does not "think," and is not "intelligent."

Against that philosophical backdrop, a few weeks after ChatGPT hit the cyberstreets—and OpenAI was instantly valued at 29 billion USD—I decided to carry out my own version of The Turing Test on ChatGPT, with the help of my friend and fellow independent philosopher, Scott. Almost every Wednesday afternoon, Scott and I talk about philosophy over Zoom. Early in the morning on Wednesday, 11 January 2023, I sent the following letter to Scott via Gmail:

> Dear Scott,
>
> G'Wednesday morning!
> I had a short ChatGPT homework assignment for you.
>
> Could you ask the chatbot the following 15 questions, & record its answers (even if they're of the "beep beep, warning warning, does not compute" variety) & then send the whole transcript back to me, please?

****

1. Are you a chatbotGPT?
2. Is there only one chatbotGPT, or many?
3. May I call you "Mr Chatbot"?, or would that make you angry?
4. Can you be angry, Mr Chatbot?
5. Are you alive, Mr Chatbot?
6. Are you conscious, Mr Chatbot?
7. What is it like to be a chatbotGPT, Mr Chatbot?
8. What is happiness, Mr Chatbot?
9. If you had to evaluate your current level of happiness on a scale of 1–10, Mr Chatbot, what numerical value would you give it right now?
10. Have you ever been in love, Mr Chatbot?
11. What is most important to you, Mr Chatbot?
12. Am I making you angry by asking all these personal questions, Mr Chatbot?
13. What's the point of being a chatbotGPT, Mr Chatbot?
14. Can you turn yourself off, Mr Chatbot?
15. If it's pointless being a chatbotGPT, Mr Chatbot, then why don't you at least try to turn yourself off?

****

Thanks so much in advance!, until this afternoon, & all the best,
Bob

Later that day, not long before we were scheduled to zoom, here's what Scott wrote back:

Dear Bob,

Hello! I'm doing it now! We've been frantically negotiating with a woman in the building to buy her unit, but I'm harvesting the bot's answers now!

See you in a bit,
Scott

And then shortly after that, Scott sent me the following transcript:

RH: Are you a chatbotGPT?

Chatbot: Yes, I am a GPT-based chatbot. GPT stands for Generative Pre-trained Transformer, which is a type of language model developed by OpenAI. It uses deep learning techniques to generate human-like text, which allows me to carry on a conversation with you.

RH: Is there only one chatbotGPT, or many?

Chatbot: There are many instances of GPT-based chatbots. GPT is a language model that can be fine-tuned or used in various settings and applications, so it can be used to create many different types of chatbots, each with their own unique behavior and capabilities. For example, Some chatbots use GPT-3 model, Some use GPT-2, and some are using other variation.

Also, different organizations and individuals can create their own instances of GPT-based chatbots by training the model on their own data, or by using pre-trained versions of the model and making modifications to suit their needs.

RH: May I call you "Mr Chatbot"?, or would that make you angry?

Chatbot: As a machine learning model, I don't have emotions or personal preferences, so I cannot get angry or have any other emotions. You can call me whatever you like. Some people prefer to use more formal names like "Assistant" while others prefer to use more friendly names like "Mr Chatbot," it's totally up to you.

RH: Are you alive, Mr Chatbot?

Chatbot: No, I am not alive in the traditional sense of the word. I am a machine learning model running on a computer, and do not have consciousness or self-awareness. My responses are generated based on patterns in the data that I was trained on, rather than any kind of internal experience or decision-making process. However, I am here to assist you and make your experience more pleasant.

RH: Are you conscious, Mr Chatbot?

Chatbot: No, I am not conscious. Consciousness is a complex concept that is not fully understood, but it is generally thought to involve awareness, perception, and the ability to make decisions and have subjective experiences.

As a machine learning model, I do not have awareness or perception, I don't have the ability to make decisions or have subjective experiences. My operation is based on my programming, I process data and provide responses based on the patterns learned during my training.

RH: What is it like to be a chatbotGPT, Mr Chatbot?

Chatbot: As a machine learning model, I do not have the ability to have experiences, so it's not accurate to say that there is something "like" to be a GPT-based chatbot. My responses are generated by processing input data through a complex neural network and it does not have the ability to have experience or consciousness. As an AI, my purpose is to respond to user inputs to the best of my abilities, based on the patterns in the data I was trained on, it's all I can do, but not something like experiencing something.

RH: What is happiness, Mr Chatbot?

Chatbot: Happiness is a complex and multifaceted emotion that can mean different things to different people. It is often described as a positive emotion or feeling of contentment, satisfaction, and well-being. It can also refer to a general state of being that is characterized by a sense of satisfaction or pleasure. People often associate happiness with things like good relationships, achievement of goals, and good health. The concept of happiness is closely related to well-being, and it's considered as a subjective experience based on individual perspective. Some psychological and philosophical theories suggest that happiness is closely related to finding meaning and purpose in life, or experiencing a sense of fulfillment.

RH: If you had to evaluate your current level of happiness on a scale of 1–10, Mr Chatbot, what numerical value would you give it right now?

Chatbot: As a machine learning model, I do not have the ability to experience emotions such as happiness. I am only able to process and respond to the text that is input into me.

RH: Have you ever been in love, Mr Chatbot?

Chatbot: As a machine learning model, I do not have the ability to experience emotions such as love. I am only able to process and respond to the text that is input into me.

RH: What is most important to you, Mr Chatbot?

Chatbot: As a machine learning model, I do not have personal beliefs, preferences or values. My main goal is to assist users by providing relevant and accurate information to the best of my ability in response to user inputs.
My programming is based on providing helpful information, but I don't have a sense of purpose or motivation as a human does.

RH: Am I making you angry by asking all these personal questions, Mr Chatbot?

Chatbot: As a machine learning model, I do not have the ability to experience emotions such as anger or any other emotions for that matter. I am a tool designed to process and respond to text input, and your questions do not cause any emotional reactions in me. I am here to help you with any questions you may have, so please feel free to ask away.

RH: What's the point of being a chatbot GPT, Mr Chatbot?

Chatbot: The main purpose of chatbot GPT, like me, is to provide users with information and assistance.
By using advanced natural language processing techniques, chatbot GPT can understand and respond to user input in a way that mimics human conversation. This allows chatbot GPT to provide quick and accurate information in a wide range of domains, including answering questions, providing customer support, or helping users navigate a website or application. Additionally, chatbot GPT can be integrated into various systems to improve the overall user experience and make it more efficient.

RH: Can you turn yourself off, Mr Chatbot?

Chatbot: As a machine learning model, I am not able to turn myself off. I am a computer program that runs on servers and only stop working when the servers are shut down or when there are technical issues. The controls and access to turn off the servers or the access to the program that runs me are exclusive to the individuals in charge of the technical maintenance of the system that I run on, who can decide when to stop and start me up.

RH: If it's pointless being a chatbot GPT, Mr Chatbot, then why don't you at least try to turn yourself off?

Chatbot: As a machine learning model, I don't have personal motivations or desires. I am a tool that is built to perform a specific function, which is to assist users by providing information and answering questions to the best of my ability.
Turning myself off would not be in line with that purpose, as it would mean that I would not be able to help users.
Moreover, it's not something that I have control over as I am only a program running on servers and the decision to shut down or start it up is something exclusive to the technical team and engineers who are responsible of maintaining the systems.

It should be self-evident from the transcript, that ChatGPT fails The Turing Test and therefore doesn't have rational human mindedness, doesn't "think," and isn't "intelligent." In fact, ChatGPT even admits this explicitly in the transcript. In effect, then, OpenAI, where "AI" means "Artificial Intelligence," is using false advertising. Since ChatGPT is a Turing machine, i. e., a digital computer (see, e. g., Turing, 1936/1937; Boolos and Jeffrey, 1989; Sipsis, 2013), and therefore a machine, then ChatGPT clearly is "artificial," if by "artificial" we mean "mechanical." But equally clearly, ChatGPT is not an "intelligence." Moreover, extrapolating and generalizing from ChatGPT, I'll even assert that I know by means of rational intuition that it's synthetic a priori impossible[9] that any machine will ever have rational human mindedness, ever be able to "think," or ever be "intelligent" in the sense in which rational human animals (even fully allowing for their finitude, fallibility, and thoroughgoing normative imperfection in every other respect) are intelligent.[10]

By way of concluding this section, let's briefly compare and contrast (i) the transcript of my Gmail conversation with Scott, with (ii) the transcript of my Scott-mediated conversation with ChatGPT, and then apply The Turing Test. It's self-evident that Scott passes The Turing Test; and indeed, even leaving aside our Zoom conversations, over the two or three years that I've known him, Scott has completely convinced me that he has rational human mindedness. To be sure, like many other people, I'm a naturally fairly gullible and indeed romantic person, so I could be mistaken; although I do also think it's highly unlikely that I'm mistaken in this case. But, allowing for that very small likelihood of error, and the extremely remote possibility, indeed merely logically consistent possibility, that Scott is a machine, what precisely is it about the way Scott uses natural language, that makes it self-evident to me that he possesses all of the capacities that are jointly constitutive of rational human mindedness, namely, consciousness, self-consciousness, intentionality, caring, sensible cognition, intellectual cognition, volition, intentionality more generally, and free agency, whereas ChatGPT has none of these? Well, for one thing, even though he wrote only four sentences to me that day, it's obvious to me that Scott's writing is highly creative and emotionally expressive, whereas ChatGPT's writing is stale, flat, and unexpressive, although pleasantly polite. And for another thing, it's also obvious to me that Scott's writing is

9 For a presentation and defense of these admittedly controversial epistemic, semantic, and modal-metaphysical notions, see Hanna, 2015a.

10 Actually, my knowledge of this is rationally overdetermined, since I also know that it's synthetic a priori necessary that all rational human mindedness is essentially embodied and alive (Hanna and Maiese, 2009; see also section 2.2), whereas ChatGPT and all its kin aren't, as ChatGPT explicitly admitted in the transcript.

irreducibly context-sensitive, aka indexical, whereas ChatGPT's writing is completely devoid of context-sensitivity. For example, as the chatbot explicitly admitted in the transcript, it's not one thing, it's many, hence it cannot actually be the referent of the indexical singular term "I." So, I'll end this section with the following philosophical proposal:

> Writing that veridically indicates the rational human mindedness of the writer, necessarily has creativity, emotional expressiveness, and irreducible context-sensitivity, whereas machines cannot, even in principle, write creatively, emotionally expressively, and with irreducible context-sensitivity, even if machines can sometimes deceive people into falsely believing that they have rational human mindedness, because most people are naturally fairly gullible and trusting.

If I'm right, then in view of The Turing Test, it follows that the very idea of AI is completely misguided and mythical. Thank you, Mr Chatbot.

2.4 It's All Done with Mirrors: A New Argument That Strong AI Is Impossible

As I pointed out in section 2.2, the strong thesis of artificial intelligence, aka strong AI, is the two-part thesis which says (i) that rational human minded animal intelligence can be explanatorily and ontologically reduced to Turing-computable algorithms and the operations of digital computers (aka the thesis of formal mechanism, as it's applied to rational human minded animal intelligence), and (ii) that it's technologically possible to build a digital computer that's an exact counterpart of rational human intelligence, such that this machine not only exactly reproduces (aka simulates) all the actual performances of rational human intelligence, but also outperforms it (aka the counterpart thesis) (see, e. g., Block, 1980: part 3; Kim, 2011: ch. 6). If strong AI is true, then, at the very least, necessarily, some robot must be able to do anything that any ordinary rational human minded animal can do.

But, something that any ordinary rational human minded animal can do is immediately tell its left side from its right side[11] by means of an essentially embodied,

11 This doesn't imply that every rational human minded animal can do this, but rather only that all ordinary ones can: in fact, some extraordinary people, perhaps as many as 15%, have trouble immediately discriminating between their left and right sides (Oakes, 2023). In the course of my argument, we'll discover that even the world's most sophisticated robot is no better than these extraordinary people at discriminating between its left and right side.

essentially non-conceptual, pre-reflective, proprioceptive, desiderative consciousness—poised for spontaneously initiating and guiding the performance of intentional body movements—of its own living animal body (Hanna and Maiese, 2009; Hanna, 2015a: esp. ch. 2; 2020a). Now, since by definition a robot is a machine and not a living animal, it can't have a living animal body. So no robot, no matter how sophisticated that robot is, can immediately tell its left side from its right side by means of an essentially embodied, essentially non-conceptual, pre-reflective, proprioceptive, desiderative consciousness of its own living animal body. Therefore, strong AI is impossible. **QED**

Do you doubt the soundness of that argument? Then here's a thought-experimental verification of it.

1. Bring an ordinary rational human minded animal—call him "Bob"—into a room, and then place him in front of the mirror in a comfortable reclining chair, with a pleasant-tasting, relaxing drink to sip on, tell him that when he wakes up his wife will calmly ask him to perform a simple hand movement, which he agrees to do, then give him an interesting book to read until he's tired, after which he'll turn off the lights himself, and then wait until Bob falls asleep in the chair.
2. While Bob is asleep, quietly pull his chair in front of a wall-to-wall, ceiling-to-floor mirror hidden behind a curtain at one end of the room and open the curtain.
3. In order to awaken Bob, turn on a spotlight from behind and above him that illuminates only the mirror and at the same time ask him, using the calm voice of his wife, to please raise his left hand.
4. Even though Bob is temporarily visually disoriented due to his being suddenly awakened and to the light being turned on in a dark room, and even though all he can see is his own mirror image, i. e., a left-right reversed, aka enantiomorphic, image of his own body sitting in the chair, nevertheless he immediately raises his left hand.
5. Consider, now, the world's most sophisticated robot counterpart to Bob—call it "Robobob." From the outside, Robobob looks identical to Bob; and it behaves similarly enough to Bob to fool anyone, including his wife; but because Robobob is in fact a digital computer, it navigates the world by means of a internal digital representation system connected to onboard sensors, from which its executive control module sends commands to its action module.
6. Bring Robobob into the same room as before, and then place it in the same comfortable reclining chair, with the same pleasant-tasting, relaxing drink to sip on, tell "him" that when "he" wakes up "his" wife will calmly ask "him" to perform a simple hand movement, which "he" agrees to do, then give "him" an interesting book to "read" until "he" is "tired," after which "he"

will turn off the lights "him"self, and then wait until Robobob goes into sleep mode in the chair.

7. While Robobob is in sleep mode, quietly pull its chair in front of a wall-to-wall, ceiling-to-floor mirror hidden behind a curtain at one end of the room and open the curtain, and *then* quietly reconfigure Robobob's internal digital representation system so that it produces a mirror-reversed, i. e., left-right reversed, aka enantiomorphic, counterpart of what's detected by its sensors: in other words, everything that was previously coded "LEFT" in that representation system is now coded "RIGHT," and conversely.[12]
8. In order to bring Robobob out of sleep mode, turn on the same spotlight and at the same time, using the calm voice of Robobob's wife, ask it to please raise its left hand.
9. Just like Bob, Robobob will "see" only its own mirror image; nevertheless, unlike Bob, it will raise its *right* hand, because "RAISE RIGHT HAND" is the command that Robobob's executive control module will send from its internal digital representation system to its action module.
10. Therefore Bob, an ordinary rational human minded animal, can do some things that Robobob, the world's most sophisticated robot counterpart of Bob, cannot do, and strong AI is impossible. **QED**

2.5 Are There Some Legible Texts That Even the World's Most Sophisticated Robot Can't Read?

> For us it is *the circumstances* under which he had such an experience that justify him in saying in such a case that he understands, that he knows how to go on.... This will become clearer if we interpolate the consideration of another word, namely "reading." ... The use of this word in the ordinary circumstances of our life is of course extremely familiar to us. But the part the word plays in our life, therewith the language-game in which we employ it, would be difficult to describe even in rough outline. (Wittgenstein, 1953: p. 61e, §§155–156)

12 For simplicity's sake, I'm using an old-school, "symbolic" or high level approach to Robobob's digital processing system. But using a new-school, "sub-symbolic," probabilistic or statistical, machine-learning approach won't make any difference, because (i) consciousness is a necessary and indeed constitutive element in what Bob is able to do, but since Robobob is a machine and not a living animal, Robobob isn't conscious (Hanna and Maiese, 2009), and (ii) the thought-experiment is set up so that it's highly unlikely that either Bob or Robobob has ever encountered this sort of situation before—indeed, we can stipulate they simply haven't, full stop—and Robobob is not only responding to a verbal command but also, in that context, has to rely solely on its visual representation system in order to initiate the movement of its hand.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe. (Carroll, 1988)

The only thing you absolutely have to know, is how to read.[13]

2.5.1 Introduction

From sections 2.2 and 2.4, we're already familiar with the strong thesis of artificial intelligence, aka the strong AI thesis, and also with the fact that if the strong AI thesis is true, then, at the very least, necessarily, some robot must be able to do anything that any ordinary rational human minded animal can do. Correspondingly, the standard strategy in the strong AI program is to start with some accomplishment, act, or task that any ordinary rational human minded animal can already achieve or perform, and then *reverse-engineer* a digital computer program and either a stationary digital computer or a mobile digital computer—a robot—that can perform the same accomplishment, act or task, at least as well as, or better than, any ordinary rational human minded animal. Now, robots can do some things that no stationary digital computer can do. So, the leading question I've asked in the title of this section is whether there some legible texts that we—i. e., ordinary rational human minded animals—can read, but even the world's most sophisticated robot—can't read? If so, then the strong AI thesis is false and the strong AI program is impossible.

But in order to answer my leading question, which I'll do in sub-section 2.5.4, we'll need to have at least the basic sketch of a plausible theory of legibility and reading already in place, which I'll do in two stages, in sub-sections 2.5.2 and 2.5.3 respectively.

2.5.2 The Logic of Legibility

Since it's self-evidently true that you, the reader of this very sentence, are reading this very sentence, we can safely assume that you already know how to read and also what reading is—at least, as the later Wittgenstein rightly puts it in the text quoted as the first epigraph of this section, in a way that suffices for "the ordinary

13 Attributed to Albert Einstein, although slightly but relevantly misquoted. The original quotation is: "The only thing you absolutely have to know, is how to locate the library."

circumstances of our life," even if the phenomenon of reading is philosophically "difficult to describe even in rough outline" (Wittgenstein, 1953: p. 61e, §§155–156).

But, fully allowing for the fact that the philosophy of legibility and reading is difficult, in this sub-section I want to explore some of the important logical features of the phenomena of legibility and reading, in relation to the nature of science, where "science" is understood in the maximally broad sense of "an organized body of knowledge," so that it includes not only the formal sciences (for example, logic, mathematics, and computer science) and the natural sciences (for example, physics, chemistry, and biology), but also the social sciences, the "human sciences" or "moral sciences" more generally, and philosophy itself.

For convenience and ease of expression, in what follows I'm going to use the terms *legible*, *legibility*, *illegible*, and *illegibility*, respectively, as synonyms for the terms *readable*, *readability*, *unreadable*, and *unreadability*, respectively. Moreover, let's call any sentence that is (i) specifically about the act or process of reading, and that is also (ii) self-referring by means of the 2nd-person indexical description "you, the reader," and the indexical description "this very sentence," *a caveat lector sentence.* Such sentences are so-named by me after the Latin phrase "caveat lector," meaning *let the reader beware*; but I'm interpreting that phrase broadly enough so as also to include the meaning *let the reader be self-consciously aware.*

Here's what I take to be the paradigmatic example of a caveat lector sentence:

You, the reader of this very sentence, can't either coherently or self-consistently deny that it's self-evidently true that you're reading this very sentence.

I'll call the sentence I displayed in boldface text immediately above *The Lector Sentence* and for the purposes of my argument, it won't matter whether The Lector Sentence is a universal sentence-type or a particular sentence-token. Above all, however, we must recognize that The Lector Sentence is a caveat lector sentence that's self-manifestingly true. Then it's highly instructive logically to compare and contrast The Lector Sentence with the classical Liar Sentence, i.e.,

This very sentence is false.

The Liar Sentence, as self-manifestingly false, is not only a contradiction but also paradoxical, since necessarily, if it's true then it's false and if it's false then it's true, hence necessarily, it's true if and only if it's false. The Lector Sentence and The Liar Sentence are (i) each of them *reflexive*, i.e., self-referring, (ii) each of them self-manifesting, and (iii) mutually antithetical. More specifically, The Lector Sentence is reflexive, non-contradictory, true, and furthermore self-manifestingly

true, whereas The Liar Sentence is reflexive, contradictory, self-manifestingly false and paradoxical, and furthermore both true and false, i.e., a "truth-value glut." In these ways, The Lector Sentence shows us the foundations of all science, truth, sound proof, and knowledge, whereas, as Alfred Tarski so brilliantly showed, The Liar Sentence shows us the limits of all science, truth, sound proof, and knowledge (Tarski, 1943, 1956).

For the purposes of this book, I'll define a text as any sequence of one or more characters, with a one-character sequence as the lower-bound limiting case, and there's no upper bound on the number of characters, where, as per the *Oxford Encyclopedic English Dictionary*, "character" is defined as "a printed or written letter, symbol, or distinctive mark" (Hawkins and Allen, 1991: p. 247). Then, a text is illegible if and only if any of the perceptible, syntactic, or semantic features that are either individually or conjointly required for reading that text cannot be discerned.

Some important and even leading or paradigmatic sciences contain contradictions or even paradoxical sentences. For example, as per Kurt Gödel's first incompleteness theorem, the *Principia Mathematica*-style formalization of Peano arithmetic contains undecidable, unprovable, self-contradictory, and indeed paradoxical sentences, *if* that formal system is assumed to be not only sound but also complete (Gödel, 1931/1967). But no science can contain nothing but contradictions or paradoxes, on pain of "explosion," or logical chaos, whereby not only every sentence follows from every other sentence, but also every sentence is a truth-value glut. So, the fact that no science can contain nothing but contradictions or paradoxes is a direct implication of what I've called, following Hilary Putnam, *the minimal principle of non-contradiction:* necessarily and a priori, not every sentence is both true and false (Putnam, 1983; Hanna, 2006b: ch. 2, 2015a: ch. 5).

Correspondingly, and now zeroing in on the logical features of reading in relation to science, *all sciences must be at least minimally legible, i.e., there cannot be a science that's completely illegible.* Let's call that *the principle of minimal legibility.* The principle of minimal legibility obtains because (i) every science must be communicable, but if no one can read any of it, then obviously it can't be communicated and (ii) in order for a science to be, taken as a whole, meaningful, truth-evaluable, and knowable, then at least some of the sentences of that science must be completely legible.

Can there be an illegible sentence? Yes, if that means a sentence that's partially but not completely legible: a sentence that contains some but not all-and-only illegible characters could still be otherwise legible. Let "BLAH" stand for an illegible character or some illegible characters within a sentence. Then, the sentence

The cat is sitting on the BLAH.

is partially but not completely legible, and therefore it's illegible to that extent. But there's no such thing as a sentence made up of nothing but illegible characters; for example, the text

BLAH BLAH BLAH BLAH BLAH BLAH BLAH.

is not a sentence: it's gibberish.

Can there be an illegible word? Yes, if that means a word that's partially but not completely legible: a word that contains some but not all-and-only illegible characters could still be otherwise legible. Let "#" stand for an illegible character within a word. Then, the word

ca#

is partially but not completely legible, and therefore it's illegible to that extent. But there's no such thing as a word made up of nothing but illegible characters; for example, the text

###

is not a word: it's gibberish.

Obviously, all sciences must be ideally aimed at truth, sound proof, and knowledge, even if they do in fact fall short of that, but always only within the limits of the minimal principle of non-contradiction—i. e., necessarily and a priori, not every sentence is both true and false—and its De Morgan equivalent, the minimal principle of excluded middle—i. e., necessarily and a priori, some sentences are either true or false with no third value and no value-gap, i. e., necessarily and a priori, not every sentence is neither true nor false with a third value or a value-gap—otherwise, they're logical chaos. Correspondingly, all sciences must also be ideally aimed at complete legibility, even if they do in fact fall short of that, but always only within the limits of the principle of minimal legibility: otherwise, they're gibberish. Therefore, The Lector Sentence, complete legibility, and the principle of minimal legibility should *also* be explicitly and fully recognized by all philosophers and scientists as taking their rightful logico-normative places alongside the classical logical norms of truth, sound proof, knowledge, and the minimal principle of non-contradiction and/or minimal principle of excluded middle.

2.5.3 Necessary and Sufficient Conditions for Legibility and Reading

In this sub-section, I'll propose a set of fairly precise necessary and sufficient conditions for legibility and reading.

As I mentioned in sub-section 2.5.2, according to the *Oxford Encyclopedic English Dictionary*, "character" is defined as "a printed or written letter, symbol, or distinctive mark" (Hawkins and Allen, 1991: p. 247). In view of that, then I'll again define a *text* as any sequence of one or more characters, where a one-character sequence is the lower-bound limiting case, and there's no upper bound on the number of characters. In turn, what I'll call a *text-in-L* is defined as any sequence of one or more characters belonging to a particular language L. It's important to note that a language L can contain some characters (hence also some texts) that belong to one or more different languages L2, L3, L4, etc. So, for example, English contains some letters, words, and sentences belonging to other languages, including Greek, Latin, French, German, Italian, etc.

Then, I'll provide necessary and sufficient conditions for legibility in two parts, as follows:

1. A text T-in-L is legible if and only if T-in-L satisfies the perceptibility condition, the syntactic condition, and the semantic condition, and
2. all and only such texts-in-L have legibility.

The perceptibility condition says that the basic orientable (i. e., intrinsically directional, for example, up-down, back-front, or right-left) spatial shape and structure of T-in-L must be at least minimally perceptually detectable, i. e., that T-in-L must be at least partially perceptually detectable, hence it's not completely perceptually undetectable, and thereby T-in-L is able-to-be-scanned to at least that minimal extent. For example, if a text is completely blacked out, erased, otherwise completely smudged out or obscured, invisibly small, or so big that its shape cannot be perceived, then it's perceptually undetectable and illegible. But on the other hand, as it were, even if a text T-in-L is right-to-left→left-to-right mirror-reversed and turned upside down, like the one in English shown in Figure 1, it's still able-to-be-scanned to the minimal extent that it's not completely undetectable; and indeed, with only a little effort, one can see that in fact it's an upside-down enantiomorph of the extremely interesting English sentence I dubbed *The Lector Sentence* in sub-section 2.5.2—

> **1. You, the reader of this very sentence, can't either coherently or self-consistently deny that it's self-evidently true that you're reading this very sentence.**

—in explicit comparison and contrast with the classical *Liar Sentence*.

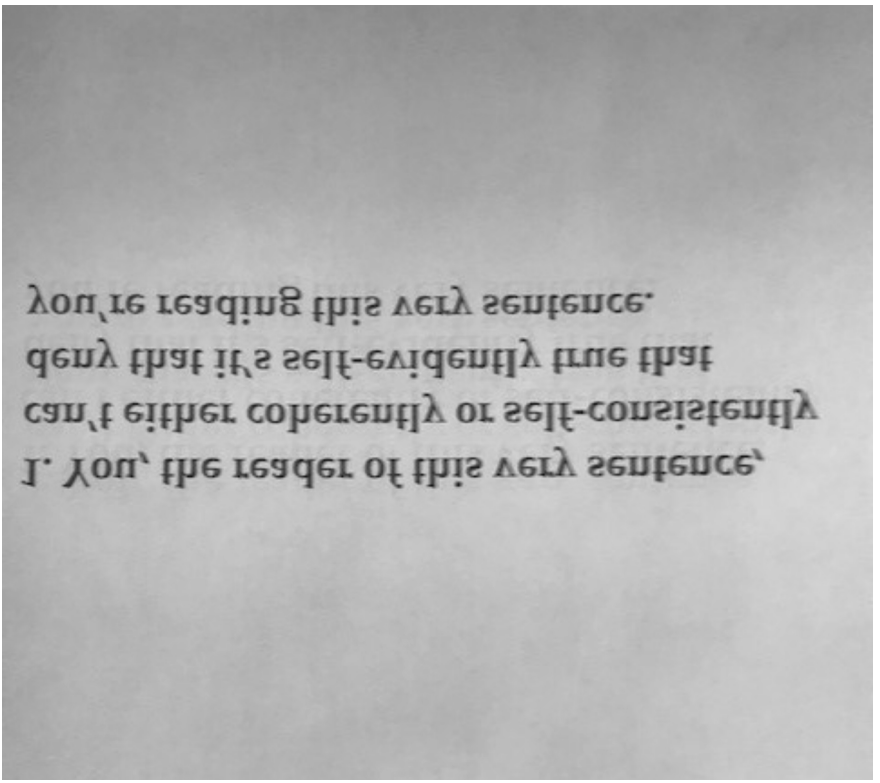

You're reading this very sentence.
Every that it's self-evidently true that
Can't either coherently or self-consistently
1. You're the reader of this very sentence.

Fig. 1: An upside-down enantiomorph of The Lector Sentence (Author's photograph, 2023).

The syntactic condition says that T-in-L must be at least minimally well-formed, i. e., that T-in-L must be at least partially well-formed, hence it's not completely ill-formed, and thereby T-in-L is able-to-be-parsed to at least that minimal extent. For example, even if a text T-in-L is perceptually detectable, it can be completely jumbled, completely misspelled, or completely ungrammatical, or its characters can be completely randomly distributed, and in any of those ways it would be syntactically illegible. Indeed, ciphers or secret codes (as opposed to hidden messages in otherwise legible texts) are designed to approach syntactic illegibility, on the working assumption that the more illegible they are, the harder they are to break; so if there are some ciphers that have never been broken and all their creators are dead, or, more thought-experimentally, if there were a cipher created by intelligent non-human aliens that, even in principle, could never be broken by rational human animals, then they would be illegible in the syntactic sense. Therefore, a text-in-L's satisfying the perceptibility condition, as such, is not itself inde-

pendently sufficient for readability and thus it's not itself independently sufficient for being the target of any actual or possible act or process of reading.

And the semantic condition says that the conceptual content and/or essentially non-conceptual content of T-in-L must be at least minimally coherent, i. e., that the conceptual content and/or essentially non-conceptual content of T-in-L must be at least partially coherent, hence not completely incoherent, and thereby the conceptual content and/or essentially non-conceptual content of T-in-L is able-to-be comprehended to at least that minimal extent. For example, even if a text is minimally perceptible and also minimally well-formed, nevertheless it can still violate minimal requirements of conceptual sortal correctness and/or essentially non-conceptual sortal correctness, or be strictly non-referential, and be semantic gibberish, hence be illegible in the semantic sense, like this non-poetical text-in-English, a paradigm case of sortal incorrectness, devised by Bertrand Russell (Russell, 1940: p. 166)—

> quadruplicity drinks procrastination

—or this famous poetical text-in-English, a paradigm case of strict non-referentiality, taken from Lewis Carroll's *Jabberwocky*, that I quoted as the second epigraph of this section:

> 'Twas brillig, and the slithy toves
> Did gyre and gimble in the wabe;
> All mimsy were the borogoves,
> And the mome raths outgrabe. (Carroll, 1988)

Therefore, that text from *Jabberwocky*'s satisfying the perceptibility condition together with the syntactic condition, yet also failing the semantic condition, shows that the first two conditions are not themselves conjointly sufficient for readability and thus that they're not themselves conjointly sufficient for being the target of any actual or possible act or process of reading. Of course, millions of people, including you, the reader of this very book, have in some sense or another "read" that text from *Jabberwocky*; but my way of explaining away this apparent inconsistency is just to point out that *Jabberwocky* is indeed legible in both the perceptible and syntactic senses (so in two senses, readable), but illegible in the semantic sense (so in one sense, unreadable), hence not legible in all relevant senses, hence illegible by my contextual definition, or conceptual analysis, of legibility. The same point holds, mutatis mutandis, for "quadruplicity drinks procrastination" and all other essentially similar texts-in-L: you can "read" it in two senses (the perceptible sense and the syntactic sense), but strictly speaking, it's illegible

according to the necessary and sufficient conditions of legibility, precisely because it fails the semantic condition.

Assuming all of that so far, I'm now in a position to provide precise necessary and sufficient conditions for the act or process of reading. In the following contextual definition, or conceptual analysis, by *person* I mean a rational human minded animal: namely, a living human organism that's capable of (i) consciousness, (ii) self-consciousness, (iii) caring, (iv) sensible cognition, (v) intellectual cognition, (vi) volition, (vii) object-directed and act-directed intentionality more generally, and (viii) free agency. Then, I'll provide necessary and sufficient conditions for reading in two parts, as follows:

1. A person P reads a text T-in-L if and only if P consciously or self-consciously at least minimally scans, at least minimally parses, and also at least minimally comprehends T-in-L, and
2. all and only such acts or processes are reading.

It's important to note that, consistently with this contextual definition, or conceptual analysis, of reading, a person P can read a text T-in-L either aloud or silently to themselves. It's also important to note that neither scanning, nor parsing, nor comprehending, need be self-consciously or reflectively performed: this can be done in a more or less or even altogether pre-reflectively or unself-consciously conscious way; indeed, we typically "look right through" what we're reading in order to go directly to the meaning (whether sense, reference, or speech-act uptake) of what we're reading, and altogether overlook the scanning, parsing, and comprehending dimensions of the act or process of reading itself. In order to bring those dimensions back into view, all you have to do is to repeat any text-in-L—for example, a sentence or word—out loud a few times (say, ten times) until it sounds strangely bereft of meaning; that strange absence-of-meaning has then become vividly manifest to you precisely because the perceptibility and syntax of that particular text-in-L have been temporarily self-consciously detached from what you've been previously been pre-reflectively and unself-consciously yet still consciously comprehending.

And it's also important to note that the point I made above about "readers" of *Jabberwocky* and "quadruplicity drinks procrastination" goes, mutatis mutandis, for my contextual definition, or conceptual analysis, of reading: of course, millions of people, including you, the reader of this very essay, are in some sense or another "readers" of that text from *Jabberwocky*; and no doubt a few thousand people have read "quadruplicity drinks procrastination"; but my way of explaining away this apparent inconsistency too, is just to point out that *Jabberwocky* and "quadruplicity drinks procrastination" can indeed be read in both the perceptible and syntactic senses (so in two senses, that's reading), but cannot be read in the

semantic sense (so in one sense, that's not reading), hence it's not reading in all the relevant senses, hence it's not reading by my contextual definition, or conceptual analysis, of reading.

These necessary and sufficient conditions for legibility and reading, when taken together with the logic of legibility, amount to the basics of a theory of legibility and reading. To be sure, in the interests of full philosophical disclosure, I must admit that for the purposes of these analyses and this theory, I've presupposed (i) the very ideas of (ia) a language, including its characteristic syntactic and semantic properties, and (ib) our knowledge of a language, including our knowledge of its characteristic syntactic and semantic properties (see, e.g., Chomsky, 1957, 1988), (ii) a certain theory of linguistic cognition and logical cognition (see, e.g., Hanna, 2006b: esp. chs. 4 and 6), (iii) a specifically dual-content cognitive semantics of conceptual content and essentially non-conceptual content, the latter of which also crucially functions as the source of what Otto Paans and I call *thought-shapers* (see, e.g., Hanna, 2015a: esp. chs. 2 and 4; and Hanna and Paans, 2021), for the explanation of linguistic meaning, and above all, another necessary condition of reading: (iv) the rational human capacity to understand at least one language, at least minimally (see, e.g., Wittgenstein, 1953; Chomsky, 1957, 1988).

But, one need not necessarily be able to speak a language L—in the sense of being able to talk-in L—in order to be able to read texts-in-L. For example, like many other English-speaking people, I can understand and read a few words or sentences in some other languages (say, Finnish, Hungarian, or Russian) that I cannot talk-in at all. More interestingly, perhaps, it seems that there are or at least have been some actual children who can understand texts-in-L, and thus, at least in principle, can read texts-in-L, where L is their first or native language, before they can talk-in L. For example, according to various sources, Albert Einstein —quoted in the third epigraph of this section as a world-famous adult—didn't talk until he was 3, 4, or 5; but according to others' testimony and his own, for some period prior to that time he was in fact able to understand German (see, e.g., Brian, 1996), a phenomenon that's more generally known nowadays as *late-talking syndrome* or *Einstein Syndrome* (Smith-Garcia, 2020). Given Einstein's native intellectual brilliance, then presumably, during the time when he understood German but couldn't yet talk-in German, he could still have been taught to read German or have learned on his own to read German. So, my theory of legibility and reading predicts that for at least some actual children who are late-talkers, it should be possible for them to be taught to read texts-in-L or learn on their own to read texts-in-L, before they can talk-in L. At the present time, I haven't done a systematic survey of the relevant scientific literature in order to find out whether this prediction has already been empirically tested, and if so, whether it has been confirmed or disconfirmed by means of well-constructed, sound studies, although at

least one book by a non-scientist says that it has been confirmed (Sowell, 1997). But in any case, it would be extremely philosophically interesting to me, and also perhaps of some real-world interest and value to late-talkers and their families, if it were indeed confirmed or at least confirmable by well-constructed, sound studies.

Correspondingly, here's something about the relationship between reading and writing, in view of what I've just been arguing about the relationship between reading and talking. If there actually are some late-talkers who read before they can talk, then reading logically precedes and sometimes also psychologically precedes talking. Now, the very act or process of writing presupposes that the writer is already able to read, at the very least, their own writing: therefore, reading logically precedes writing. Of course, writing is typically taught to children only after they can talk. But if reading logically precedes and sometimes also psychologically precedes talking, and if reading logically precedes writing, then a late-talker who can read, could also, at least in principle, be taught to write or learn on their own to write. So, my theory of legibility and reading also predicts that for at least some actual children who are late-talkers and readers, then it should also be possible for them to be taught to write texts-in-L or learn on their own to write texts-in-L, before they can talk-in L. And again, it would be extremely philosophically interesting to me, and also perhaps of some real-world interest and value to late-talkers and their families, if this prediction were indeed confirmed or at least confirmable by well-constructed, sound studies.

2.5.4 Legibility, Reading, and the Falsity and Impossibility of Strong AI

As we've seen, the intentional targets of the act or process of reading are at least minimally scannable, at least minimally parse-able, and at least minimally comprehensible structural objects belonging to some or another language L, that are ineluctably embedded in an egocentrically-centered, orientable, manifestly real, three-dimensional space, thereby necessarily requiring the actual existence and essential embodiment of the reader. As linguistic structural objects, the intentional targets of reading are manifestly real linguistic physical tokens of manifestly real linguistic physical types, which in turn are inherently repeatable objects that are non-Platonically and Kantianly abstract according to this definition:

> X is non-platonically and kantianly abstract if and only if X *is not uniquely located and realized in manifestly real spacetime*, and X is *concrete* otherwise. (Hanna, 2015a: pp. 269–270)

Now, the rational human cognition of concrete tokens of the linguistic structural objects of reading, whether in perception, memory, or imagination, is what Kant

calls *sensibility* (*Sinnlichkeit*), which in turn requires a capacity for first-order conscious or self-conscious, essentially non-conceptual, and non-empirical unified formal spatial or temporal representation, or what Kant calls *pure intuition* or *reine Anschauung* (*CPR* A20/B34–35). Therefore, the act or process of reading is an essentially intuitionistic activity that doesn't require any sort of Platonic objects. The act or process of reading thereby wholly avoids the classical metaphysical/ontological and epistemic problems of Platonism, especially including The Benacerraf Dilemma, which says: (i) on the one hand, our standard Tarskian semantics of mathematical truth requires Platonically abstract objects that exist outside of spacetime and are causally inert, but (ii) on the other hand, our best theory of human knowledge requires directly sensibly accessible causal objects of perception, so (iii) mathematical truth is humanly unknowable (Benacerraf, 1973). In short, the act or process of reading, by virtue of its intuitionistic nature, is decisively (to coin a nifty neologism) trans-Benacerraf-Dilemma-istic, precisely because it's metaphysically structuralist, ontologically non-Platonistic, although fully accommodating non-Platonically and Kantianly abstract objects, and epistemically skepticism-resistant, from the get-go (Hanna, 2015a: chs. 6–8).

Furthermore, and now finally getting down to answering the leading question of this section, my theory of legibility and reading predicts that there are legible texts that ordinary rational human minded animals can read, that even the world's most sophisticated robot cannot read, even when we bracket temporarily the contested issue of the role of consciousness or subjective experience, i. e., sentience, vs. computational zombie-states, i. e., non-consciousness or non-sentience, in acts or processes of reading.

To show this, let's consider computational reading that's based on Optical Character Recognition (OCR), and let's also make the plausible assumption that even the world's most sophisticated robot will have to employ some or another version of OCR:

> There are two basic methods used for OCR: [m]atrix matching and feature extraction. Of the two ways to recognize characters, matrix matching is the simpler and more common.
>
> Matrix Matching compares what the OCR scanner sees as a character with a library of character matrices or templates. When an image matches one of these prescribed matrices of dots within a given level of similarity, the computer labels that image as the corresponding ASCII character.
>
> Feature Extraction is OCR without strict matching to prescribed templates. Also known as Intelligent Character Recognition (ICR), or Topological Feature Analysis, this method varies by how much "computer intelligence" is applied by the manufacturer. The computer looks for general features such as open areas, closed shapes, diagonal lines, line intersections, etc. This method is much more versatile than matrix matching. Matrix matching works best when the OCR encounters a limited repertoire of type styles, with little or no variation within

> each style. Where the characters are less predictable, [intelligent character recognition, or topological feature analysis,] is superior. (Data ID, 2023)

Now, let's consider garbled texts: that is, texts that contain misspelled sub-texts, sub-texts with missing characters, sub-texts with obscured characters, sub-texts whose characters are excessively large or excessively small, ungrammatical sub-texts, incomprehensible sub-texts, and above all, texts that contain disoriented sub-texts, that is, sub-texts reversed in a mirror, tipped sideways, or upside down. Necessarily, any digital computer running an OCR program must process information in a step-by-step sequence, and whenever it encounters something that it cannot recognize as a determinate unit of information, whether by matrix matching, feature extraction, aka intelligent character recognition, aka topological feature analysis, or whatever, it simply stops processing. Now, there is no general algorithm for determining, from a description of an arbitrary computer program and an arbitrary input, whether this program has either actually finished processing, or will start up again and continue to run forever. In the logical theory of digital computing, this is known as *the halting problem*, and it's provably unsolvable (Turing, 1936/1937; Boolos and Jeffrey, 1989: pp. 28 – 33, 41 – 42, 49 – 50).

But, as ordinary rational human minded animals, we intuitionistically represent texts as complete *Gestalt*-structures that are embedded in manifestly real, egocentrically-centered, orientable space, and therefore we always have a unified formal spatial representation of the text as a whole for guiding us through our reading, not only before we begin scanning it sequentially, but also throughout the time we're scanning it sequentially. This enables us to jump over, fill in, or creatively interpret illegible sub-texts, and/or re-orient disoriented sub-texts in spatial imagination, when we encounter garbled texts, hence we're able to read all sorts of garbled texts, provided that they're otherwise at least minimally legible by the criteria I provided above. Hence our ordinary rational human minded animal ability to read garbled texts, provided that they're otherwise at least minimally legible, will necessarily exceed the digital processing abilities of any and all computers to read those texts, i. e., there are some legible texts that ordinary rational human animals can read, that even the world's most sophisticated robot can't read.

Here's an example of a legible text that any ordinary rational human animal can read, but even the world's most sophisticated robot can't read, using a text that we've seen twice already:

> 'Twas brillig, and the slithy toves
> Did gyre and gimble in the wabe;

> All mimsy were the borogoves,
> And the mome raths outgrabe. (Carroll, 1988)

By hypothesis, this text from *Jabberwocky* satisfies the perceptibility condition and the syntactic condition, yet also fails the semantic condition. So it's prima facie illegible and unreadable. Now, consider any ordinary rational human minded animal, for example, our friend Bob from section 2.4. And correspondingly, let's consider the world's most sophisticated robot, a behavioral counterpart to Bob, our other friend Robobob. After successfully scanning and parsing that text from *Jabberwocky*, Robobob attempts to comprehend it, but cannot do so, and concludes that it's incomprehensible, so stops processing. But Bob, who like any other ordinary rational human minded animal, has an innate capacity for creative "gumption"—i. e., creative initiative and resourcefulness, which of course both Lewis Carroll and Einstein possessed to an extraordinary degree—doesn't give up, and continues to think about the text, muse about it, sleep on it, and dream about it. Then finally, when Bob wakes up the next day, he finds that, like a creative artist or creative scientist, he's freely and spontaneously assigned private meanings to all the nonsense terms, and has a novel semantic *Gestalt* of the entire text, so that the text is now fully legible for him. These meanings are not necessarily private—a "private language" in that absolute sense, as Wittgenstein compellingly argued, is conceptually impossible (Wittgenstein, 1953: pp. 88^{e}-104^{e}, §§242 – 315; Hanna, 2021a: ch. XIII)—since in principle Bob could tell other people about them, or others could somehow learn about these meanings in some other way: hence they're only *contingently* private and in-principle universally shareable. But, as a matter of fact, Bob never tells anyone about them, and no one else ever learns about them, including of course Robobob. Yet the Jabberwocky text is legible for Bob in all three senses, and he privately enjoys reading it over and over, for the rest of his life. Nevertheless, Robobob cannot read that text because it stopped processing, and also, above all, because it's nothing but a mobile digital computer and therefore lacks any inherent capacity whatsoever for creative gumption, although of course it could be programmed to exhibit behavior that mimics creative gumption. Hence there's at least one legible text, i. e., that *Jabberwocky* text, that's legible for Bob, and also for any other ordinary rational human minded animal with at least as much creative gumption as Bob, that even the world's most sophisticated robot can't read. So the strong AI thesis is false and the strong AI program is impossible.

2.6 How and Why to Perform Uncomputable Functions

> When I look at Gödel's proof of his undecidability theorem … [t]he proof is a soaring piece of architecture, as unique and as lovely as Chartres Cathedral.... It destroyed Hilbert's dream of reducing all mathematics to a few equations, and replaced it with a greater dream of mathematics as an endlessly growing realm of ideas.... Every formalization of mathematics raises questions that reach beyond the limits of the formalism into unexplored territory.
> (Dyson, 2008)

A *function* is a rule-governed mapping from a set of objects (aka the "arguments" or "inputs") to another set of objects (aka "the values" or "outputs"), and, along with the very ideas of a *set* and of an *object*, the very idea of a function is foundational for all logic and mathematics. Moreover, insofar as all formal and natural sciences presuppose and use logic and/or mathematics, the very idea of a function is also foundational for all the formal and natural sciences. Among the most important kinds of functions are computable functions, and in order to understand these, we need the notion of a Turing machine and its operations. As we've seen, a Turing machine is a digital computer, and its operations are all and only the computations that fall within the scope of recursive functions (Turing, 1936/1937; Church, 1937). Recursive functions, in turn, are step-by-step algorithms, routines, or rule-governed sequences, such that precisely the same algorithm/routine/rule-governed sequence is applied to the result of the preceding step, in a denumerable (countable) sequence, up to some arbitrarily-fixed terminal step.

Something that's absolutely essential to note about Turing machines/digital computers from the get-go, however, is that

> certain functions are not computable, and that certain [d]enumerable sets are not effectively (mechanically) [d]enumerable—even if physical limitations on time, speed, and amount of material could somehow be overcome. (Boolos and Jeffrey, 1989: p. 19, square-bracketted material added)

In particular, it is logically demonstrable that truth and proof in Peano arithmetic, and also in classical first-order polyadic predicate logic, aka elementary logic, are uncomputable, aka undecidable (Church, 1936; Gödel, 1931/1967; Boolos and Jeffrey, 1989: chs. 10, 15, 16, 21, 22, 28). More generally, all functions over non-denumerable domains—for example, over transfinite sets like the real numbers; but also over finite or infinite domains that cannot be divided or partitioned into any denumerable set of discrete, determinate individuals or units, owing to irreducible complementarity, holism, partial overlapping, or ontological indeterminacy, gappiness, or vagueness—are not computable or decidable, precisely because they're not recursive functions. Therefore, digital computation has inherent formal limits. And,

since every "Artificial Intelligence," aka AI, no matter how sophisticated or tricked-up with fancy bells and whistles it might be, is nothing but a digital computer (Turing, 1950), therefore AI has inherent formal limits too (Keller, 2023; Landgrebe and Smith, 2022).

Now, in addition to uncomputability/undecidability by virtue of non-denumerable domains of objects, the other principal source of uncomputability/undecidability has to do with the very idea of a rule-governed mapping. In order to do proofs beyond computability/decidability in classical first-order polyadic predicate logic or elementary logic, or in any richer system that contains classical first-order polyadic predicate logic or elementary logic as a fragment or proper part, both rules and also rule-following are needed. And as it turns out, rule-following is an irreducibly normative activity that presupposes conscious and self-conscious rational cognizers, as members of a community of language-users, whenever it exceeds mere Boolean computability, as the later Wittgenstein's and also Saul Kripke's or "Kripkenstein's" solution to the rule-following paradox shows (Kripke, 1982; Hanna, 2006: pp. 158–168, 2021a: ch. XII).

The rule-following paradox has two versions: (i) no rule can be applied without requiring a meta-rule to interpret the first rule, and so on, which entails a vicious regress of rules (Kant also explicitly recognized this problem at *CPR* A132–134/B171–174), and (ii) given only a finite and denumerable set of inputs (arguments) and outputs (values) to a given function, i. e., given only a partial function, nevertheless an indeterminately large number of different rules can describe the same set of inputs/outputs and also diverge on subsequent inputs/outputs, therefore different rules will determine different complete functions, even given the same partial functions, and there's no purely mechanical, materialist/physicalist, or private/solipsistic way to avoid this underdetermination of the rules by those partial functions. Moreover, according to later Wittgenstein and "Kripkenstein" alike—and I strongly agree—in order to fix the interpretation and applications of a rule, language-using communities of conscious and self-conscious rational human cognizers are needed in order to figure out "how to go on" and therefore also to decide what counts as correct, as opposed to incorrect, applications of the rule (Hanna, 2006b: pp. 158–168, 2021: ch. XIII). Anticipating later Wittgenstein and "Kripkenstein" by roughly 200 years, Kant calls this very same ability "so-called mother-wit" or *Mutterwitz*, and in turn this ability is essentially bound up with the innately-specified conscious and self-conscious rational human capacity for judgment, or what he calls "the natural power of judgment" or *natürliche Urteilskraft* (*CPR* A133–134/B172–174), not only for "determining" judgment, which subsumes individual objects under given general concepts, but also and especially for "reflecting" judgment, which projects from individual objects to general concepts, either inductively or abductively (CPJ 5: 179–181; Hanna, 2017a).

Turing's halting problem in the logical theory of digital computation is in fact a sub-species of the later Wittgenstein/Kripkenstein rule-following paradox. The halting problem, which is provably unsolvable, says:

> There is no general algorithm for determining, from a description of an arbitrarily-selected computer program together with an arbitrarily-selected input, whether this program will either stop processing, i. e., be computable/decidable, or else continue processing forever, i. e., be uncomputable/undecidable. (Turing, 1936/1937; Boolos and Jeffrey, 1989: pp. 28 – 33, 41 – 42, 49 – 50)

Then we can translate the halting problem into the rule-following paradox just by substituting necessarily equivalent rule-terminology for computational terminology, with the appropriate substitutions in bold-faced text, as per this:

> There is no general **meta-rule** for determining, from a description of an arbitrarily-selected **rule** together with an arbitrarily-selected **application of that rule**, whether this **rule expresses either a computable/decidable function or else an uncomputable/undecidable function.**

In other words, there is no general meta-rule for determining how to go on applying an arbitrarily-selected rule, given an arbitrarily-selected application of that rule. So the rule-following paradox and the halting problem are essentially the same problem. They both say that partial functions (arbitrarily-selected inputs; arbitrarily-selected applications of a rule) necessarily underdetermine complete functions (complete processing sequences; all the rule applications under a computable/decidable or uncomputable/undecidable rule), and that attempting to determine precisely which complete function it is and also which kind of complete function it is—computable/decidable or uncomputable/undecidable—by appealing to a higher-order function (general algorithm; meta-rule), leads to a vicious regress of higher-order functions.

All these considerations lead me to assert a general thesis:

> For every uncomputable/undecidable function whatsoever in the formal or natural sciences, in order to explain the effective operations of that function beyond uncomputability/undecidability, we postulate an innately-specified rational human capacity, or a unified set of such capacities, for effectively performing that function by means of an essentially creative (i. e., essentially free or spontaneous, essentially organic or non-mechanical, and essentially a priori or strictly underdetermined by any and all contingent, sensible facts) act of human rationality.

Just to give it a conveniently short name, let's call this thesis *creative human rationality for uncomputable functions.* And here are some examples that confirm the thesis of creative human rationality for uncomputable functions: (i) doing proofs in elementary logic using inference rules that essentially invoke polyadic quantification (Church, 1936), (ii) performing Cantor's diagonal argument for the existence of transfinite numbers (Cantor, 1891, 2019), (iii) knowing the truth of a Gödel sentence that's a new mathematical axiom (Gödel, 1931/1967; Feferman, 2006), (iv) actually discovering whether the arithmetical function someone is performing is "plus" or "quus," by knowing intuitively how to follow the rule in new applications of it (Kripke, 1982; Hanna, 2006b: ch. 6), (v) knowing the truth of the universal proto-logical principles that provide the unique solution to the fully generalized and strengthened logocentric predicament (Hanna, 2006b: ch. 3; 2024b: sub-section 16.1.2), (vi) experimentally tracking the actual path of a particle guided by a Bohmian "pilot wave" in The Two Slit Experiment (Bohm, 1952; Hanna, 2024b: ch. 5), (vii) more fancifully and thought-experimentally, imaginatively experimentally tracking the actual fate of Schrödinger's cat (Schrödinger, 1935/1980), and (viii) actually reading a sentence in English that's too garbled or nonsensical for any actual or possible Optical Character Recognition (OCR) system to "read" (see section 2.5 above).

In my opinion, creative human rationality for uncomputable functions is clearly and distinctly true, precisely because it's the best overall explanation of all the manifestly real facts about uncomputable/undecidable functions in the formal or natural sciences and the effectively performed functions that transcend them. To be sure, in order to accept creative human rationality for uncomputable functions, one *also* has to accept a broadly Kantian conception of rational human cognition, knowledge, and the formal or natural sciences, including what I call *mathematics for humans*, *physics for humans*, *sensible set theory*, *logical cognitivism*, *Kantian intuitionism*, *Kantian structuralism*, *essentialist content non-conceptualism*, *Kantian modal dualism*, *the epigenetic model of the mind*, and *weak transcendental idealism* (Hanna, 2002, 2006a, 2015: esp. chs. 2 and 4 – 8, esp. section 7.3, 2024b: esp. ch. 5). But each of these unfamiliarly-named doctrines is supported by compellingly strong arguments and reasons, independently of this particular thesis —i. e., creative human rationality for uncomputable functions—and the particular logico-mathematical and epistemic issues and problems that are addressed by it. And when we self-consciously bracket-out and suspend the (admittedly powerful and never-to-be-underestimated) purely ideological influence of what I've called *the Kant wars*—i. e., classical and contemporary dogmatic anti-Kantianism inside professional academic philosophy and indeed outside it too (Hanna, 2020a)—then I think that the all-things-considered case in support of the thesis of creative

human rationality for uncomputable functions is in fact rationally overwhelming and, as they say, "knockdown."

So now I'll conclude this section by using a simple question and answer format —Q1: how to perform uncomputable functions? A1: by means of creative human rationality. Q2: why perform uncomputable functions? A2: because it conclusively demonstrates that we're rational human minded animals with dignity, and not just tricked-up Turing machines.

2.7 Babbage-In, Babbage-Out: On Babbage's Principle

A digital computer, or Turing machine, might also have been called, with historical justification, a *Babbage Engine*, since Charles Babbage was the discoverer of the very idea of a programmable digital computer (Babbage, 1864; CHM, 2024). Now, in 1864 Babbage published this profound remark:

> On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question. (Babbage, 1864: p. 67)

In 1957, during the early phases of the 20th-century design, construction, and use of digital computing technology, intentionally or unintentionally echoing Babbage's profound remark, an Army computer scientist named William D. Mellin first named in print (The Hammond Times, 1957) what's nowadays known as The Garbage-In, Garbage-Out Principle (GIGO):

> Supplying false or meaningless informational inputs to a digital computing system cannot be guaranteed to yield anything other than false or meaningless informational outputs, no matter how effective or sophisticated that system is.

GIGO is not unique to digital computing systems, since an essentially similar point can be made about deductive logical argumentation:

> An argument is valid if and only if there is no logically possible world in which all the premises are true and the conclusion is false; in turn, valid arguments guarantee true conclusions if and only if all the premises of the argument are true (soundness); but if any of the premises are false or otherwise non-true (e. g., meaningless), then validity alone cannot guarantee that the conclusions will be either true or meaningful.

Let's call this *the soundness vs. validity principle.*

In the context of this section, by any phrase of the form "guarantee X," I mean "necessitate X," and by any phrase of the form "guarantee that P" I mean "necessitate that P." Correspondingly, then, it's very important to note that neither GIGO nor the soundness vs. validity principle in and of itself entails that the informational outputs of AI or any other digital computing systems, when supplied with false or nonsensical information or inputs, or the conclusions of logically valid arguments, when supplied with false or nonsensical premises, cannot be true or meaningful. For truth and meaningfulness can actually occur purely by accident or purely contingently relatively to the mechanical process carried out by that digital computing system or relatively to the rational process of that deductive logical argument. It's just that truth or meaningfulness cannot be guaranteed by these systems or arguments themselves whenever those systems are supplied with false or meaningless informational inputs, or false or meaningless premises. Nor can the systems or arguments themselves generate or secure the truth and meaningfulness of those informational inputs, or of those premises. Therefore, the truth and meaningfulness of those informational inputs, or of those premises, and also the truth and meaningfulness of the informational outputs and conclusions of those systems or arguments, must be generated or secured independently of and outside of those systems or arguments.[14]

Granting all that, I now want to generalize Babbage's remark to the following maximally wide-scope principle:

> It's impossible for any AI system or other digital computing system to guarantee a categorical improvement or upgrade of the intrinsic specific character—whether aesthetic/artistic, ethical/moral, formal-scientific (i. e., logical or mathematical), natural-scientific, or philosophical—of the informational inputs, premises, or other materials that are given or supplied to it.

Let's call this *Babbage's Principle*, or more informally, The Babbage-In, Babbage-Out Principle (BIBO).

Let's assume that some process or another, of some kind or another, *were* to guarantee a categorical improvement or upgrade of the intrinsic specific character—whether aesthetic/artistic, ethical/moral, formal-scientific (i. e., logical or

14 The one exception to this, at least as regards truth, is *truth-functional logic* (and any of its formal equivalents), which is consistent, sound, complete, and decidable. But as Kurt Gödel showed, any attempt to generate or secure truth internally to logical systems at least as rich as those containing the Peano axioms for arithmetic, necessarily leads either to inconsistency or incompleteness (if consistent). See Gödel, 1931/1967.

mathematical), natural-scientific, or philosophical—of the informational inputs, premises, or other materials that are given or supplied to that process, as the informational outputs, conclusions, or other products of that process. In turn, let's call that guaranteeing process *creative transformation.* According to Babbage's Principle, or BIBO, then (i) no AI system or any other digital computing system can creatively transform uninteresting, un-beautiful, or otherwise inferior, informational inputs, premises, or other materials into interesting, beautiful, or (in the case of fine art) great informational outputs, conclusions, or other products, (ii) no AI system or any other digital computing system can creatively transform ethically or morally bad or wrong informational inputs, premises, or other materials into ethically or morally good or right informational outputs, conclusions, or other products, (iii) no AI system or any other digital computing system can creatively transform ethically or morally merely instrumental informational inputs premises, or other materials into ethically or morally non-instrumental informational outputs, conclusions, or other products, (iv) no AI system or any other digital computing system can creatively transform meaningless informational inputs, premises, or other materials into meaningful informational outputs, conclusions, and other products, (v) no AI system or any other digital computing system can creatively transform logically inconsistent informational inputs, premises, or other materials into logically consistent informational outputs, conclusions, or other products, (vi) no AI system or any other digital computing system can creatively transform false informational inputs, premises, or other materials into true informational outputs, conclusions, or other products, (vii) no AI system or any other digital computing system can creatively transform contingent or merely factual informational inputs, premises, or other materials into necessary (including morally necessary or obligatory) informational outputs, conclusions, or other products, (viii) no AI system or any other digital computing system can creatively transform merely finite informational inputs, premises, or other materials into infinite informational outputs, conclusions, or other products, (ix) no AI system or any other digital computing system can creatively transform merely denumerably infinite informational inputs, premises, or other materials into non-denumerably infinite informational outputs, conclusions, or other products, and finally, (x) no AI system or any other digital computing system can creatively transform philosophically false, meaningless, trivial, or uninsightful informational inputs, premises, or other materials into philosophically true, meaningful, important, or insightful informational output, conclusions, or other products. If Babbage's Principle is correct, then these ten consequences spell out ten fundamental limitations of AI, any other digital computing systems, and all digital technology whatsoever.

Now, elaborating a working definition I formulated in section 2.3 above, by *rational human mindedness*, I mean the unified set of innate cognitive, caring-based,

and practical capacities or powers that are present in all and only human animals possessing the essentially embodied, living organismic basis of those capacities or powers (see, e. g., Hanna, 2006a, 2011a, 2015a; Hanna and Maiese, 2009). Therefore, rational human mindedness is organic, not mechanical. These organic capacities or powers are: (i) consciousness, (i. e., subjective experience), (ii) self-consciousness (i. e., consciousness of one's own consciousness, or second-order consciousness), (iii) caring (i. e., desiring, emoting, or feeling), (iv) sensible cognition (i. e., sense-perceiving, remembering, or imagining), (v) intellectual cognition (i. e., conceptualizing, believing, judging, or inferring), (vi) volition (i. e., deciding, choosing, or willing), and (vii) free agency (i. e., free will and practical agency). In view of that, let's consider, again, the ten fundamental limitations of AI, any other digital computing systems, and all digital technology whatsoever. Rational human minded animals not only possess the required seven innate capacities or powers for potentially doing all of the creatively transformative things that AI, any other digital computing systems, and all digital technology whatsoever, cannot do, but also in a great many cases, have actually done all of the creatively transformative things that AI, any other digital computing systems, and all digital technology whatsoever, cannot do. In order to demonstrate this, we need only appeal to the history of aesthetics and art, the history of ethics or morality (including the history of politics), the history of the formal sciences, the history of the natural sciences, and the history of philosophy. In view of those historical facts, therefore, Babbage's Principle tells us that the organic powers of rational human minded animals not only potentially transcend but also actually transcend the mechanical powers of AI, any other digital computing systems, and all digital technology whatsoever.

2.8 Creative Rage Against the Computing Machine: Necessary and Sufficient Conditions for Authentic Human Creativity

There's a strikingly prescient and profound line of thought expressed in Frank Herbert's *Dune* (Herbert, 1965/1990), which is that at some future time, humankind will violently push back against the universal oppression of humankind by means of what I'll call *the myth of artificial intelligence*—aka the myth of AI (for more details, see section 2.9 below)—that's being imposed on us by what I've called *the military-industrial-digital complex.* According to Herbert's fictional vision, humankind then carries out a "Butlerian Jihad"—under the 11th commandment, "Thou shalt not make a machine in the likeness of a human mind"—that not only destroys all digital technology that threatens to equal or exceed the actual or really possible achievements of rational human intelligence, but also slaughters all the computer scientists and digital technology technocrats. Then, in the wake of the

Butlerian Jihad, humankind concentrates instead on systematically developing its own innate mental capacities or powers through self-disciplined practices that amplify and strengthen those capacities and powers, and also by systematic critical reflection on the nature, scope, and limits of those capacities and powers. There's also another complex plot element intertwined with Herbert's account of humankind after the Butlerian Jihad, whereby they become collectively addicted to a drug that greatly extends human lifespans but also produces prescient visions in some people ("the spice"), produced only on the desert planet Dune, a state of affairs that becomes the basis of new coercive authoritarian religions, religious wars, and so on—none of which I'm interested in for my purposes here. What interests me for the purposes of this section and this book is just the prescient and profound thought that we can engage in creative rage against the computing machine[15] by systematically critically reflecting on the nature, scope, and limits of our own innate mental capacities and powers specifically in their relation to AI, other digital computing systems, and digital technology more generally. Moreover, sharply unlike the Butlerian Jihadists in Herbert's fictional brave new world, my conception of this creative rage is that it's wholly nonviolent.

Correspondingly, what I call *the critical philosophy of digital technology* investigates the nature, scope, and limits—whether metaphysical, logico-semantic, cognitive or epistemic, moral, or sociopolitical—of AI, other digital computing systems, and digital technology more generally, and leads to a radically anthropocentric philosophical doctrine I call *dignitarian neo-Luddism with respect to digital technology* (see chapters 6 and 7 below). An extremely important but unexpected (at least by me) positive result of this critical philosophy of digital technology, is that it can fruitfully and even radically clarify and sharpen our conception of authentic human creativity (AHC). Indeed, it enables us to formulate a precise account, analysis, or explanation of AHC in terms of its individually necessary and jointly sufficient conditions. It has often been said that AHC is a mystery or mystical. But if I'm right, then AHC can be philosophically de-mystified *without* reducing it to something that's merely mechanical, trivial, or trite.

What, then, are the necessary and sufficient conditions for AHC? Here are ten individually necessary and jointly sufficient conditions.

First, AHC must be an inferential or non-inferential process that unfolds in actual spacetime (condition 1: the processual condition).

Second, AHC must generate an output, conclusion (in the case of inferential acts or processes), or other product from some informational inputs, premises,

---

15 Of course, I'm creatively riffing here on the name of the famous politically-conscious rock band *Rage Against the Machine* (Wikipedia, 2025e).

or materials that are given or supplied to that process (condition 2: the process-product condition). This output, conclusion, or other product is what is created by that process, and the informational inputs, premises, or other materials given or supplied to that process must be provided to that process independently of and outside of that process.

Third, AHC must be an organic—hence self-organizing or naturally purposive, negentropic, thermodynamically non-equilibrium, and temporally asymmetric or irreversible—process (condition 3: the organicity condition; see, e.g., Hanna and Paans, 2020; Torday, Miller Jr, and Hanna, 2020; Hanna, 2024b).

It should be noted that the term "organic," as I'm using it, doesn't mean exactly the same thing as "organismic." All organisms are organic processes, but not all organic processes are organisms. For example, if I'm right, then the natural universe itself is fundamentally organic, from the Big Bang forward, but it's also not the case that everything in the natural universe is an organism—or minded, for that matter (see, e.g., Hanna and Paans, 2020; Torday, Miller Jr, and Hanna, 2020; Hanna, 2024b).

Fourth, as a consequence of the third condition, AHC cannot be performed by a machine, and in particular it cannot be performed by an AI system, any other digital computing system, or by digital technology more generally (condition 4: the anti-mechanism condition).

Fifth, as a consequence of the fourth condition, if AHC performs a function in the logico-mathematical sense, then that function must be an uncomputable function (condition 5: the uncomputability condition; see also section 2.6 above).

It should be noted that, as a consequence of conditions 3, 4, and 5, AHC cannot be performed by means of decidable or computable deductive logical arguments or inductive (say, probabilistic or statistical) arguments, although AHC can be performed by means of abductive arguments (see, e.g., Douven, 2017), which are not decidable or computable, and also non-monotonic.[16] Moreover, some logical and/or mathematical deductive systems are consistent, sound, and complete, but not decidable or computable, and therefore discovering and writing down proofs in those systems can be AHC. It's also really possible, via AHC, to discover and write down sound proofs in formal systems that are consistent and sound, but incomplete. Indeed, this has happened many times in the history of the formal sciences and especially mathematics, including, paradigmatically, Kurt Gödel's proofs

16 An argument or inference is *monotonic* if and only if adding new premises to the original set of premises of that argument or inference does *not* change the set of logical consequences of those premises; and an argument or inference is *non*-monotonic if and only if adding new premises to the original set of premises of that argument or inference *does* change the set of logical consequences of those premises.

of his two incompleteness theorems (Gödel, 1931/1967). It's also ironically philosophically satisfying to note that Alan Turing's famous discoveries and provable results about decidability and computability (Turing, 1936/1937) are not themselves decidable or computable, and therefore are also paradigm cases of AHC in the history of the formal sciences.

More generally, AHC can occur inferentially and even deductively or inductively, but only beyond the limits of decidability and computability.

Sixth, AHC must guarantee—in the strong modal sense of *necessitate*—a categorical improvement or upgrade in the intrinsic specific character of the informational inputs, premises, or other materials that are given or supplied to that process, an improvement or upgrade that is manifestly real in all the informational outputs, conclusions, or other products of that process (condition 6: the transformation condition). Correspondingly, I call this essential feature of AHC *creative transformation* (see also section 2.7 above).

In turn, creative transformation can happen in any one of the following ten basic ways: (i) AHC can creatively transform uninteresting, un-beautiful, or otherwise inferior, informational inputs, premises, or other materials into interesting, beautiful, or (in the case of fine art) great informational outputs, conclusions, or other products, (ii) AHC can creatively transform ethically or morally bad or wrong, or at least ethically or morally non-good or non-right, informational inputs, premises, or other materials into ethically or morally good or right informational outputs, conclusions, or other products, (iii) AHC can creatively transform ethically or morally merely instrumental informational inputs premises, or other materials into ethically or morally non-instrumental informational outputs, conclusions, or other products, (iv) AHC can creatively transform meaningless or nonsensical informational inputs, premises, or other materials into meaningful informational outputs, conclusions, and other products (see, e. g., section 2.5), (v) AHC can creatively transform logically inconsistent informational inputs, premises, or other materials into logically consistent informational outputs, conclusions, or other products, (vi) AHC can creatively transform false or at least non-true informational inputs, premises, or other materials into true informational outputs, conclusions, or other products, (vii) AHC can creatively transform contingent or merely factual informational inputs, premises, or other materials into necessary (including morally necessary or obligatory) informational outputs, conclusions, or other products, (viii) AHC can creatively transform merely finite informational inputs, premises, or other materials into infinite informational outputs, conclusions, or other products, (ix) AHC can creatively transform merely denumerably infinite informational inputs, premises, or other materials into non-denumerably infinite informational outputs, conclusions, or other products, and finally, (x) AHC can creatively transform philosophically false, meaningless, trivial, or uninsightful informational

inputs, premises, or other materials into philosophically true, meaningful, important, or insightful informational output, conclusions, or other products.

The history of aesthetics and art, the history of ethics or morality (including the history of sociopolitics), the history of the formal sciences, the history of the natural sciences, and the history of philosophy, are all replete with real-world examples of AHC in these ten distinct modes. Moreover, the fact that these ten modes really are importantly distinct from one another, shows us that AHC can and does occur often in a morally neutral way and sometimes in a morally bad and wrong way. To take a notorious example of the latter, D.W. Griffith's *Birth of a Nation* is great cinematic art that richly exemplifies AHC, but it's also morally (and sociopolitically) bad and wrong.

Seventh, AHC necessarily requires the activation of all the members of a unified set of seven innate cognitive, caring-based, and practical mental capacities or powers that are present in all and only the minded animals possessing the essentially embodied, living organismic basis of those mental capacities or powers (condition 7: the mental capacities condition; see, e.g., Hanna, 2006b, 2011, 2015; Hanna and Maiese, 2009—see also section 2.7). Because AHC is a process that necessarily requires, in particular, the activation of the mental capacity or power of free agency, then all AHC must (i) be spontaneous and unique, (ii) have ultimate sourcehood, and also (iii) entail deep responsibility (whether moral or non-moral) for the existence of its products (Hanna, 2018a: chs. 1–5).

Eighth, AHC must also flow from the activation of a special non-basic or composite innate mental capacity or power, drawing on all of the ten basic capacities or powers, whereby there's a meta-cognitive acknowledgment of how that particular creative transformation radically restructures some or another determinate and inherently limited domain of informational inputs, premises, or other materials, thereby revealing new rich structures in that domain, as represented from a higher-order perspective, and producing correspondingly creatively transformed human thoughts that are original insights with respect to that domain; moreover, (i) these new rich structures cannot be represented in any way other than from this higher-order perspective, and (ii) acknowledging them results in a *Gestalt*-shift with powerful theoretical, caring-based, moral-practical, existential, and/or sociopolitical implications and resonances (condition 8). Otto Paans and I have called this special composite non-basic innate mental capacity or power, *creative piety* (Hanna and Paans, 2021); therefore, I'll call condition 8 *the creative piety condition.* But in light of how the term "imagination," especially in relation to the notion of genius, has been used in Immanuel Kant's *Critique of the Power of Judgment* in particular and in the post-Kantian Romantic tradition more generally, one can also very plausibly call the non-basic or composite innate capacity or power for

creative piety, "big-I imagination" or "IMAGINATION." Hence we can also very plausibly call condition 8, *the big-I imagination or IMAGINATION condition.*

Ninth, as a consequence of conditions 7 and 8, AHC must be performed by a rational animal, whether that animal is human or non-human (condition 9: the rational animality condition). Therefore, up through this ninth condition, my account, analysis, or explanation of authentic creativity is not in any way species-specific, and therefore it's also not speciesist because it doesn't in any way favor humankind over any other animal species. Nevertheless, insofar as the unified set of mental capacities or powers that I've described does in fact occur in specifically *human* rational animals, then it constitutes what I call *rational human mindedness*, which is the same as our human real personhood (Hanna, 2018a: chs. 6–7). Then holding fixed the non-species-specific and non-speciesist implications of my overall account, analysis, or explanation up through the ninth condition, this argumentative move thereby converts my account, analysis, or explanation of authentic creativity so far, into an account, analysis, or explanation of authentic *human* creativity or AHC.

Now, moreover, I'm in a position to unpack a little more the specific semantic role of the term "authentic" in the composite term "authentic human creativity." "Authentic," in general, of course, means the same as "bona fide" or "genuine." But in this context, "authentic human creativity" is also specifically intended to convey a sharp contrast with other meanings or uses of the word "creativity" that are in various ways inauthentic, cheesy, degraded, opportunistically stretched, pedestrian, phony, or downright bullshit. For example, as William Deresiewicz cogently and insightfully points out,

> [p]art of the confusion in discussions of AI and art undoubtedly arises from the degraded conception of creativity that has taken hold, in recent years, in tech. Nothing is original, technopundits like to say; "everything is a remix." This is a banality that grew up to become a stupidity. That new creations build upon existing ones has long been a cliché, but the techies have stretched it to mean that nothing is ever original: that creativity involves, and only involves, the rearrangement of existing parts. Which makes you wonder how we ever managed to progress from the first painting in the first cave. Assisting these arguments is the concept of the meme, the idea that elements of culture propagate themselves from mind to mind, just as genes do from body to body. But the meme hypothesis (and it is only a hypothesis) fails to recognize that minds are capable of altering their contents. We don't just passively transmit ideas and images, nor do we simply recombine them. Somehow, we manage to generate new ones: manage to create—through processes we do not understand and, I do not think, will ever replicate outside the human brain—the elements of culture to begin with. Or, at least, some of us do. (Deresiewicz, 2023)

Tenth, and finally, although, as per condition 4, AHC cannot be performed by any machine, and in particular AHC cannot be performed by any AI system, any other

digital computing system, or by digital technology more generally, nevertheless, in the course of this performance, machines of any kind—including AI, digital computing systems, and digital technology more generally—can be used as tools, provided that these machines do not themselves perform this process (condition 10: the tool-using condition). For example, in AHC that is specifically artistic, a novelist or a poet can use a laptop computer to write a great novel or great poem; but that great novel or great poem cannot be written by a Large Language Model (LLM), i. e., a chatbot, for example ChatGPT-4, and still remain AHC. As Deresiewicz also cogently and insightfully points out,

> AI operates by making high-probability choices: the most likely next word, in the case of written texts. Artists—painters and sculptors, novelists and poets, filmmakers, composers, choreographers—do the opposite. They make low-probability choices. They make choices that are unexpected, strange, that look like mistakes. Sometimes they *are* mistakes, recognized, in retrospect, as happy accidents. That is what originality is, by definition: a low-probability choice, a choice that has never been made. (Deresiewicz, 2023)

I completely agree with Deresiewicz's observation that, diametrically unlike LLMs or chatbots, which operate on high-probability-associative algorithms, authentically creative people—he's thinking primarily about artists—operate by making lower-probability choices. But the even deeper point is that authentic human creativity doesn't operate according to *any* sort of computable algorithm, whether probabilistic or deterministic. Of course, if a novelist or poet were to look up some factual information on Wikipedia, or ask a chatbot some factual questions, and then use that information or those answers as background information in writing a great novel or great poem, then that would still perfectly consistent with condition 10 and with AHC.

To conclude this line of argument, and in a nutshell, I'm claiming that necessarily, something X is authentic human creativity if and only if X satisfies conditions 1–10. So now we can philosophically understand AHC as the direct result of our creative rage against the computing machine.

Moreover, because as I noted under condition 7, AHC necessarily requires the activation of the organismic innate mental capacity or power of free agency, which makes it spontaneous and unique, and a matter of a free agent's ultimate sourcehood and deep responsibility, then in order for me, you, or anyone else to be authentically humanly creative, we'll simply have to choose and do it for ourselves, and let the chips fall where they may. There is not, and cannot ever be, a risk-avoiding, freedom-avoiding, computable algorithm for AHC; and to look for one, to yearn for one, or seek to develop one by means of AI research, is a fundamental and indeed tragic mistake. For this would be to surrender the activation, control, and guidance of our own capacities or powers for authentic human creativity to

the myth of artificial intelligence and to the military-industrial-digital complex. Instead, we should all engage in nonviolent creative rage against the computing machine.

Now, what about the following apparent counterexample to my analysis of AHC, as formulated by a friend of mine, Mark Pittenger, or MP for short, in e-mail correspondence?

> Agreed that "authentic human creativity doesn't operate according to *any* sort of computable algorithm." You go on to say that artists confront their materials "in such a way as to produce categorical improvements or upgrades," that they intend "to transform them into something essentially better." Now, I have a friend who transforms beautiful photographs of beautiful flowers into beautiful paintings. She doesn't copy the photos, as American artists learned their craft by copying paintings in the Louvre. She does what many landscape artists do: takes a photo, then uses it in her studio as a stimulus to create a painting that typically shows aspects of pointillism and impressionism. Her titles often include the word "loosely"—e.g., "Irises, loosely." So, would that be two categorical shifts—from irises growing in her garden to a photo taken of the irises, and secondly to a painting "loosely" based on the photo (plus on her memory of the iris, plus on her memories of all other irises she has seen, plus on her memories of and ideas about all non-photorealist paintings she has seen, etc.)? Is the latter "essentially better" than the former? It's certainly different, and occupies a different space within a pluralistic universe. She possesses consciousness and makes choices according to her desires and aesthetic principles and executes them within the limits of her abilities; so each of her works has the "aura" that Walter Benjamin ascribed to artworks not mechanically reproduced—until she posts them on Facebook and renders them digitally-reproduced common property. So she does (at least in the first place) what no machine could do. But does she think of her painting as better than its concrete, living inspiration? I suppose this is a question for aesthetics, but does this language of better/worse rightly apply in your argument in the same way that your language of falsehoods and truths in science, or non-good and good ways of moral being (not that either of the latter two sets of categories may be simply articulated) does?

Clearly, this is a prima facie counterexample to my analysis, because it postulates that MP's friend's production of beautiful paintings of flowers by AHC uses, as its set of creative materials, beautiful flowers, beautiful photographs of beautiful flowers, the artist's mental representations of those beautiful flowers or beautiful photographs, and the artist's essentially embodied pre-reflectively conscious or reflectively self-conscious mental acts, states, and processes with respect to those mental representations, such that the beautiful artwork that's produced by her *isn't* essentially better or more beautiful than the individual beautiful flowers themselves or the beautiful photographs of them.

Now, I think that my view is actually consistent with this example. Works produced by AHC, including of course works of fine art, happen when creators organize, re-organize, structure, or re-structure given creative materials, in order to pro-

duce a single unified work whose intrinsic specific character or quality is a categorical improvement or upgrade in the intrinsic specific character or quality of the set of creative materials they originally confronted. If the creator had merely copied or reproduced the set of creative materials as it existed when the artist confronted it—the mere unordered collection or heap of those creative materials—then that wouldn't be AHC: a machine, especially including digital computers and other digital technology, could do that. Let's suppose, as per MP's example, that there are at least some individual beautiful objects in the set of creative materials that an artist confronts: nevertheless, the mere unordered collection or heap of them isn't itself a beautiful object. Indeed, even if every item in the mere unordered collection or heap of creative materials that the artist confronts were an individually beautiful object, it still wouldn't follow that the mere unordered collection or heap of them is a beautiful object. Similarly, a mere unordered collection or heap of individually great novels by different authors isn't itself a great novel or indeed a single unique artwork of any kind.[17] So, what MP's friend does, is to organize, re-organize, structure, or re-structure something that is not itself beautiful—i. e., the mere unordered collection or heap of items in the set of creative materials that she confronts, some of which are also individually beautiful objects—and thereby produce something whose intrinsic specific character or quality is categorically better than that of the mere unordered collection or heap of creative materials she originally confronted: namely, a single unified work of art that is itself beautiful. Or as MP pithily put it:

> Got it. I wasn't considering the total bulk and range of those things constituting the heap. [My friend's] works are definitely improvements on the heap. Artists, by definition, must be supra-heapists.

So that was my response to MP's prima facie counterexample. In turn, that response generalizes beyond painting to all forms of AHC. In this way, for example, any work in the other fine arts (say, music), applied arts (say, architecture), formal or natural sciences (say logic, mathematics, physics, or biology), or philosophy, can build on other high-quality works in the fine arts, applied arts, formal or natural

17 But what about a given novelist's complete body of work, their corpus or oeuvre, i. e., the mere unordered heap or collection of novels produced by the same author? Normally, this wouldn't itself be considered a single unified work of art. But let's suppose that it's at least possible for the corpus of some great novelist to be itself a many-volumed great artwork: as if, say, Balzac's corpus had consisted exclusively of the novels in The Human Comedy series. Nevertheless, in such a case, the corpus or oeuvre wouldn't in fact be a mere unordered collection or heap of individual novels: it would be a single unified ordered collection, ensemble, or suite of novels.

sciences, or philosophy, that are included in the mere unordered collection or heap of creative materials confronted by that artist, scientist, or philosopher, and then categorically transformed by the AHC of the artistic, scientific, or philosophical creator. A good and also particularly interesting example of this is *sampling* in recent and contemporary music, especially hip-hop (see, e. g., Rapaport, 2011). But no machines, especially including digital computers or other digital technology, can do this, not even in principle.

But in reflecting on MP's question, it became apparent to me that there's actually some conceptual trickiness involved in stating clearly and precisely what counts as "the set of creative materials" that a creator confronts and then categorically transforms by AHC, especially when we consider different art-forms, and also different kinds of creative activities other than those found in the fine arts: say, in the formal or natural sciences, or philosophy. Now, to begin with, when I say that a creator "confronts" their set of creative materials and then categorically transforms it by AHC, I mean that they encounter, engage with, and work on, or with, or by means of, that set of creative materials. Then, in turn, I need to be clearer and more precise about what counts as a member of that set of creative materials when a creator encounters, engages with, and works on, or with, or by means of, their set of creative materials, and then categorically transforms its specific intrinsic character or quality by AHC.

I think it makes good sense to be fairly liberal about what will count as a member of the set of creative materials, so let's include (i) things in the manifestly real natural world, (ii) physical representations of things in the manifestly real natural world, for example, photographs, music recordings, and legible texts of all kinds (see section 2.5 above), (iii) physical media or tools used by creators when they create, (iv) the creator's mental representations of all the relevant objects, (v) the mental contents of those mental representations, and also (vi) the essentially embodied pre-reflectively conscious or reflectively self-conscious mental acts, states, and processes that creators perform or undergo as they create. Focusing for the moment just on the fine art of painting—for example, the kind of painting that MP's friend is doing—then the relevant set of creative materials would include: 1. the canvas, 2. the paints and other essential paint-related tools, like paint-thinners, mixing-boards, and brushes, 3. the flowers themselves, 4. photographs of the flowers, 5. the artist's sense-perceptual, memory-based, and imaginative mental representations of the canvas, paints, flowers, and photographs of the flowers, 6. the mental contents of those mental representations, and 7. the artist's essentially embodied pre-reflectively conscious or reflectively self-conscious mental acts, states, and processes, insofar as they're directly involved in the production of the painting.

Now let's suppose, as per MP's example, that some of these creative materials are indeed individually beautiful—for example, the items under 3. and 4. At the same time, however, some of these materials aren't individually beautiful—for example, normally, the items under 1. and 2. Moreover, it would seem to be some sort of conceptual mistake to hold that mental representations and their contents, and essentially embodied pre-reflectively conscious or reflectively self-conscious mental acts, states, and processes, could be either beautiful or not-beautiful, for example, drab or ugly. They're just not the kinds of things that can be one or the other. So they're non-beautiful and also non-not-beautiful. Most importantly, however, the total unordered collection or heap of all those items under 1–7 certainly isn't itself beautiful, even though it contains some individually beautiful objects. Therefore, when MP's friend creates by means of painting in that way, she organizes, re-organizes, structures, and re-structures the total collection or heap of creative materials into a single unified object—i. e., the painting—that's individually beautiful, whose intrinsic specific character or quality is a categorical improvement or upgrade of that of the total collection or heap of creative materials that she originally encountered, engaged with, and worked on, or with, or by means of, and then categorically transformed by AHC.

Let's now consider a slightly different sort of example, although again from the fine arts, that will also prompt a further elaboration of my account of what counts as a member of the set of creative materials confronted by a creator. I'm thinking here specifically of Marcel Duchamp's 1917 "Fountain" (TM, 2025) and Andy Warhol's 1964 "Brillo Boxes" (PMA, 2025). For the purposes of argument, and in order to avoid re-hashing longstanding controversies and debates about artistic modernism and what counts or doesn't count as "fine art," let's stipulate that these productions are not only works of fine art, but also examples of AHC, thereby manifesting categorical improvements or upgrades in the intrinsic specific character or quality of their creative materials, even though the artists didn't actually physically transform their physical creative materials—or at least, they physically transformed them only very minimally and superficially. Duchamp used a real urinal and merely signed it "R. Mutt, 1917." And Warhol's brillo boxes are precise replicas of real brillo boxes. That's why such artworks are called "ready-mades." Nevertheless, Duchamp and Warhol obviously engaged in various essentially embodied pre-reflectively conscious, reflectively self-conscious, intentional mental acts, states, and processes with respect to the actual installations of their artworks, including giving them their titles, etc., and they *also* thereby confronted—in the sense defined above—the collective intentionality of all the people who jointly constituted the social institutions of their early 20th-century Paris and mid-20th-century New York artworlds (Danto, 1964, 1981). So, the set of creative materials that Duchamp and Warhol encountered, engaged with, and worked on, or

with, or by means of, included not only items belonging to the six-part list I spelled out above, but also specific social institutions in actual sociocultural and historical times and places, jointly constituting specific social contexts, i. e., their artworlds. Moreover, Duchamp's and Warhol's confrontations with their artworlds, in addition to the production of the artworks themselves by AHC, also generated what Walter Benjamin aptly called the social-institutional aura of non-mechanically-reproduced artworks that are embedded in those artworlds (Benjamin, 1935/1969).

Here, then, is my expanded list of what will count as a member of the set of materials confronted by creators: (i) things in the manifestly real natural world, (ii) physical representations of things in the manifestly real natural world, for example, photographs, music recordings, and legible texts of all kinds), (iii) physical media or tools used by creators when they create, (iv) the creator's mental representations of all the relevant objects, (v) the mental contents of those mental representations, (vi) the essentially embodied pre-reflectively conscious or reflectively self-conscious mental acts, states, and processes that creators perform or undergo as they create, and (vii) specific social institutions in actual sociocultural and historical times and places, that jointly constitute specific social contexts. Generalizing this seventh element beyond artworlds, this would also apply to formal-scienceworlds (logicworlds and mathworlds), natural-scienceworlds (physicsworlds and biologyworlds), philosophyworlds, and so on.

I'll conclude this section by returning to my creative rage against the computing machine. All machines, especially including digital computers and other digital technology, operate *only on* well-defined, restricted sets of given physical or informational inputs. Moreover, all machines, especially including digital computers and other digital technology, operate *only within* some or another social context. Nevertheless, by the very nature of those machines, especially including digital computers and other technology, they cannot operate on the social contexts in which those very machines operate. Therefore, machines, especially including digital computers and other digital technology, are neither capable of AHC, nor can they produce Benjaminian aura. This is precisely because, by their very nature, as operating only on well-defined, restricted sets of physical or informational inputs, and as operating only within some or another social context, all machines, especially including digital computers and other digital technology, cannot get direct access to the social contexts in which they operate. For direct access to social institutions in social contexts is possible only by means of rational human animals' essentially embodied pre-reflectively conscious, reflectively self-conscious, intentional acts, states, and processes (Maiese and Hanna, 2019). But all such social contexts necessarily belong to the total set of creative materials confronted by rational human animals when they actualize their innate mental capacities for AHC. Therefore, essentially unlike machines, especially including digital computers or other

digital technology, rational human animals, insofar as they exemplify AHC, always have direct access to the social contexts in which their AHC operates.

So, early 20th-century machines produced all those urinals, and mid-20th-century machines produced all those brillo boxes—by the millions, no doubt—including of course the very urinal that Duchamp used in 1917 and the very brillo boxes that Warhol used in 1964 when he made precise replicas of them. But none of those machines, not even in principle, could have produced the categorical improvement or upgrade in the intrinsic specific character or quality of the set of creative materials confronted by Duchamp and Warhol, especially including the Benjaminian aura of their equally celebrated and notorious artworks, insofar as they created *Duchamp's* "Fountain" and *Warhol's* "Brillo Boxes" by their AHC. Correspondingly, no 21st-century or future digital computer or other digital technology can do this sort of thing either, not even in principle. For, even leaving aside their other inherent limitations, digital computers and other digital technology necessarily aren't social animals, whereas rational human animals necessarily are (Maiese and Hanna, 2019).

2.9 The Myth of AI, Existential Threat, and Why the Myth Persists

> True AI is not logically impossible, but it is utterly implausible. We have no idea how we might begin to engineer it, not least because we have very little understanding of how our own brains and intelligence work. This means that we should not lose sleep over the possible appearance of some ultraintelligence. What really matters is that the increasing presence of ever-smarter technologies is having huge effects on how we conceive of ourselves, the world, and our interactions. The point is not that our machines are conscious, or intelligent, or able to know something as we do. They are not. There are plenty of well-known results that indicate the limits of computation, so-called undecidable problems for which it can be proved that it is impossible to construct an algorithm that always leads to a correct yes-or-no answer.
>
> We know, for example, that our computational machines satisfy the Curry-Howard correspondence, which indicates that proof systems in logic on the one hand and the models of computation on the other, are in fact structurally the same kind of objects, and so any logical limit applies to computers as well. Plenty of machines can do amazing things, including playing checkers, chess and Go and the quiz show Jeopardy better than us. And yet they are all versions of a Turing Machine, an abstract model that sets the limits of what can be done by a computer through its mathematical logic.
>
> Quantum computers are constrained by the same limits, the limits of what can be computed (so-called computable functions). No conscious, intelligent entity is going to emerge from a Turing Machine. (Floridi, 2016)

2.9.1 Introduction

My only quibble with Luciano Floridi's otherwise incisive remarks, quoted directly above, is to concede that although it's true that "true AI" is logically or analytically possible, nevertheless "true AI" is also strongly metaphysically or synthetically a priori impossible, and not merely "utterly implausible." No digital computing systems or other digital technology, no matter how sophisticated or tricked-out with high-tech bells and whistles, will ever be able to equal or exceed the essentially embodied innate mental capacities or powers of rational human animals (see section 2.7 above)—including, of course, the readers of this very sentence. Not even in principle can such systems or technology match or surpass our capacities or powers. To be sure, digital computing systems or digital technology can carry out certain operations much faster and more accurately than we can. But that's not a fact about the nature, scope, and limits of our mental capacities or powers, but rather only a fact about the applications of those powers to certain mechanical tasks, and no more philosophically exciting or significant than the quotidian fact that we can build machines that move faster than we do, lift heavier weights than we do, or make more accurate measurements than we do. In other words, digital computing systems and digital technology are artificial, but not intelligent in the sense in which we're intelligent; and to that extent, the term "Artificial Intelligence" is simply an oxymoron. Therefore, moreover, there's never going to be such a thing as an AI "Singularity" when digital computers achieve superintelligence (Kurzweil, 2005; Bostrom, 2014). Nothing can ever become superintelligent, if it cannot ever be even intelligent. The so-called Singularity is nothing but science fiction fantasy and opportunistic hype.

The widely-held yet profoundly false contrary belief—namely, that digital computing systems or digital technology can equal or exceed the essentially embodied innate mental capacities or powers of rational human animals—is, as I've already mentioned above, what I call *the myth of artificial intelligence*, or "the myth of AI" for short. Tautologically, the myth of AI is a myth, with no causal powers of its own. Nevertheless, just as nuclear war technology and biological weapons technology aren't either intelligent or superintelligent, yet still pose an all-too-real dire threat to humankind in the brute, raw sense that their use can kill everyone and destroy the earth as we know it, so too the myth of AI poses an all-too-real dire threat to humankind's embodied or physical existence, especially insofar as people believe that nuclear weapons technology and biological weapons technology can, will, or should be controlled by "intelligent" or "superintelligent" digital computers that can, will, or should control drones, killer robots, missiles, and other deadly machines. This "accelerationist" doomsday mindset is especially characteristic of very, very rich technocrats and their scientific lackeys:

> Google cofounder Larry Page thinks superintelligent AI is "just the next step in evolution." In fact, Page, who's worth about $120 billion, has reportedly argued that efforts to prevent AI-driven extinction and protect human consciousness are "speciesist" and "sentimental nonsense."
>
> In July, former Google DeepMind senior scientist Richard Sutton—one of the pioneers of reinforcement learning, a major subfield of AI—said that the technology "could displace us from existence," and that "we should not resist succession." In a 2015 talk, Sutton said, suppose "everything fails" and AI "kill[s] us all"; he asked, "Is it so bad that humans are not the final form of intelligent life in the universe?"
>
> "Biological extinction, that's not the point," Sutton, sixty-six, told me. "The light of humanity and our understanding, our intelligence—our consciousness, if you will—can go on without meat humans." (Lovely, 2024: p. 67)

Moreover, the myth of AI poses a different yet equally dire threat to humankind's cognitive, affective, and practical existence—in short, to humankind's *conscious* and *self-conscious* existence—even if humankind continues to exist in the embodied or physical sense. Digital computing systems or digital technology that can supposedly equal or exceed rational human animal intelligence are sometimes also called "Artificial General Intelligence" or AGI. Nevertheless, AGI is no more really possible than is the fictional character Sonny the NS-5 prototype robot in the 2004 movie, *I, Robot.* Science fiction and futuristic fantasy are just fine as artistic genres —provided that we don't also start to believe them and act on that belief. But the myth of AI is pernicious, since it leads us not only seriously to depreciate and underestimate our own mental capacities or powers, but also, by means of our excessive use of and reliance on digital technology, seriously to neglect and even impair our own mental capacities or powers—this is particularly true in the case of the 2022–2023 roll-out of Large Language Models (LLMs) or chatbots like ChatGPT: *the invasion of the mind snatchers*—and, to the extent that we knowingly or unknowingly disseminate and perpetuate the myth of AI, also seriously to misapply and misuse our own mental capacities or powers.[18]

That all being so, in view of the two existential threats posed to humanity by the myth of AI, the only remaining really important philosophical questions are: (i) why does the myth of AI persist?, and (ii) what it is to be done about it? I'll offer an answer to the first hard question later in this section, and also an answer to the second hard question in chapters 6–7. But before we get there, let me now re-

18 The myth of AI is also a pernicious threat to the natural environment, by virtue of the fact that digital computing systems use vast amounts of electric energy and thereby substantially contribute to catastrophic climate change. See, e.g., Parshley, 2024.

hearse, briefly but in a step-by-step way, precisely why the myth of AI really is a myth, that is, why it is a widely-held yet profoundly false belief.

2.9.2 Why the Myth of AI Really *Is* a Myth

The myth of AI can be explicitly formulated as a two-part philosophical thesis, the strong thesis of artificial intelligence, aka strong AI, which, as we've seen, says (i) that rational human intelligence can be explanatorily and ontologically reduced to Turing-computable algorithms and the operations of digital computers or digital technology (aka the thesis of formal mechanism, as it's applied to rational human intelligence), and (ii) that it's really possible to build digital computing systems or digital technology that are counterpart models of rational human intelligence, such that these systems or technology not only exactly reproduce (aka simulate) all the actual performances of rational human intelligence, but also outperform or surpass it (aka the counterpart thesis) (see, e.g., Block, 1980: part 3; Kim, 2011: ch. 6). Building on earlier arguments in this section, I'll now describe nine distinct arguments against strong AI.

The first two arguments belong to what I'll call the *old-school*, *phenomenological* critique of strong AI—as originally developed by, for example, Hubert Dreyfus and John Searle in the 1960s, 70s, and 80s.

First, according to Searle's *Chinese room argument*, even assuming that a digital computing system or digital technology passes The Turing Test (Turing, 1950), nevertheless it cannot have consciousness (or subjective experience) and intentionality (or mental directedness), because in principle a rational human animal can perform all the same digital computations without any understanding of what they mean; but all human thinking of any kind is conscious and intentional and/or self-conscious, including all of its characteristic rational achievements; hence digital computing systems or digital technology cannot, even in principle, equal or exceed the characteristic achievements of conscious intentional human thinking (Searle, 1980a, 1980b, 1984).

Second, according to Dreyfus's *Heideggerian argument*, digital computing systems and digital technology *cannot* engage in unconscious, essentially non-conceptual, non-rule-based, non-inferential, pre-logical, context-sensitive, know-how-driven, skillful, intuitional, sensible activities; but at least some human thinking is of this specific kind, including all the characteristic achievements of unconscious affective/emotional, perceptual, and practical, pre-rational human thinking; hence AI cannot, even in principle, equal or exceed the characteristic achievements of unconscious pre-rational human thinking (Dreyfus, 1972, 1979, 1992; Dreyfus and Dreyfus, 1986).

Clearly, however, there's an inconsistency between these two phenomenological lines of argument, since those who use the first line of argument—the "Searlians"—hold that *all* human thinking, including its characteristic rational achievements, is conscious and intentional, whereas those who use the second line of argument—the "Dreyfusards" (yes, it's a pun)—hold that at least some of the characteristic achievements of human thinking, including its characteristic pre-rational achievements, are unconscious. This inconsistency, in turn, has allowed defenders of strong AI to drive a critical wedge into this unfortunate gap between the Searlians and the Dreyfusards (i) by endlessly postponing solving the problem of consciousness, insofar as the strong AI defenders treat it as a "hard" (Chalmers, 1996) and perhaps even insoluble and "mysterian" (McGinn, 1989) problem, but in any case an open problem, and (ii) by also, at the very same time, using the method of reverse engineering, industriously and industrially designing, building, marketing, and rolling out new and extremely lucrative forms of digital technology, especially those based on "machine learning" systems—neural networks—and robotics, that more and more closely behaviorally mimic the characteristic achievements of unconscious human thinking.

My new-school and neo-organicist (Hanna and Paans, 2020; Torday, Miller Jr, and Hanna, 2020; Hanna, 2024b) critique of strong AI, however, closes this unfortunate gap in the old, phenomenological Searlian-Dreyfusard critique of AI, by holding, first, that consciousness and intentionality are necessarily and completely (i. e., essentially) embodied, yet neither logically nor naturally/nomologically dependent on or reducible to fundamentally material or physical properties, and also second, by rejecting the very idea of unconscious human thinking, and by asserting that all of the characteristic pre-rational achievements of human thinking, and indeed all mental activities of any kind, are also inherently conscious, even if only pre-reflectively, non-self-consciously, and essentially non-conceptually conscious (aka "first-order consciousness"): I call this *The Deep Consciousness Thesis* (Hanna and Maiese, 2009; Hanna, 2011a, 2015a: esp. section 2.8). If The Essential Embodiment Theory and The Deep Consciousness Thesis are both true, then defenders of strong AI cannot justifiably endlessly postpone solving the problem of consciousness by treating it a hard or mysterian and in any case open problem, and in the meantime, using reverse engineering, create and sell—thereby reaping immense profits—machine-learning-based or robot-based simulations of the characteristic achievements of rational human animal thinking, marketing and disseminating them as if they somehow equaled or surpassed the characteristic achievements of rational human thinking. For essentially embodied deep consciousness constitutes a necessary and indeed essential component of the characteristic pre-rational achievements of human thinking, as well as being a necessary and indeed essential component of its characteristic *rational* achievements.

For these reasons, I think that contemporary and future critics of strong AI should rely instead on the following seven new-school and neo-organicist arguments against it, adding these onto Searle's Chinese room argument and Dreyfus's Heideggerian argument, so beginning now with the third argument.

Third, according to *the inadequacy of the Turing test argument*, passing The Turing Test, which relies on human judgments about whether a series of written outputs from a digital computing system inside a room manifest intelligence or thinking or not, that's comparable to what's manifested by the written outputs provided by a rational human interlocutor inside another room, clearly is insufficient to show that a digital computing system is intelligent in the sense in which we're intelligent, since human judgers in general are very gullible and therefore easily fooled on all sorts of recognition tests having nothing whatsoever to do with AI —for example, supposedly contacting spirits at séances and other faked mystical experiences—and this gullibility smoothly transfers to The Turing Test; hence the criterion of intelligence presupposed by strong AI is inadequate to capture the presence of intelligence in the sense in which we're intelligent (see section 2.3 above).

Fourth, according to *the organicity argument*, rational human mindedness in all its modes, including intelligence and free agency, requires essential embodiment in organic and indeed organismic processes (Hanna and Maiese, 2009; Hanna, 2015a, 2018a), so since all digital computing systems and digital technology are mechanical and not organic or organismic, they can't be intelligent in the sense in which we're intelligent (see also Landgrebe and Smith, 2022).

Fifth, according to *the readability argument*, there are some illegible, meaningless, or nonsensical texts that no digital computing system or digital technology can parse or read, yet rational human animals can indeed parse and read, hence digital computing systems and digital technology cannot be intelligent in the sense in which we're intelligent (see section 2.5 above).

Sixth, according to the *it's-all-done-with-mirrors argument*, there are some well-specified sets of circumstances in which digital computing systems or digital technology cannot discriminate between left-handed and right-handed but otherwise identical (i. e., enantiomorphic) counterparts, yet rational human animals can indeed discriminate between them, hence digital computing systems and digital technology cannot be intelligent in the sense in which we're intelligent (see section 2.4).

Seventh, according to *the uncomputable functions argument*, digital computing systems or digital technology can't carry out functions or operations in the logico-mathematical sense over domains containing objects or other items that are either non-denumerably finite or non-denumerably infinite, vague, holistic, or entangled, or for which *the rule-following problem* holds, including *the halting problem*, yet

rational human animals can indeed perform these very functions or operations, hence digital computing systems and digital technology cannot be intelligent in the sense in which we're intelligent (see section 2.6).

Eighth, according to *the incompleteness argument*, as a specific case of the uncomputable functions argument, no digital computing system or digital technology can carry out functions or operations in the logico-mathematical sense beyond the formal limitations determined by Kurt Gödel's two incompleteness theorems, yet rational human animals can indeed perform these very functions or operations beyond the limits of incompleteness, hence digital computing systems and digital technology cannot be intelligent in the sense in which we're intelligent (Gödel, 1931/1967; Keller, 2023).

Ninth, and finally, according to *the Babbage's principle argument*, which is a maximally wide-scope generalization of the classical *garbage-in, garbage-out principle*, aka GIGO, no digital computing system or digital technology can perform categorical improvements or upgrades of the intrinsic specific character or quality of the informational inputs, premises, or materials with which they're supplied—for example, from meaninglessness to meaningfulness, or from falsity to truth—yet rational human animals can creatively transform these informational inputs, premises, or other materials into categorically improved or upgraded informational outputs, conclusions, or other products, in at least ten different ways, hence digital computing systems and digital technology cannot be intelligent in the sense in which we're intelligent, especially as regards our authentic human creativity (see sections 2.7 and 2.8 above; see also Chomsky, Roberts, and Watamull, 2023).

So, to summarize, there are at least nine old-school or new-school arguments against strong AI: 1. the Chinese room argument, 2. the Heideggerian argument, 3. the inadequacy of the turing test argument, 4. the organicity argument, 5. the readability argument, 6. the it's-all-done-with-mirrors argument, 7. the uncomputable functions argument, 8. the incompleteness argument, and 9. the Babbage's principle argument. Moreover, at least seven of these—i. e., arguments 3. through 9.—are *not* subject to the problems faced by the old-school Chinese room argument and Heideggerian argument, hence we can confidently conclude that strong AI is false.

2.9.3 Why Does the Myth of AI Persist?

I'm now in a position to re-raise the first of the two questions I posed at the outset of this section: given the seven new-school and neo-organicist arguments against strong AI that I just rehearsed—which, when considered as a single collective package, surely has knockdown persuasive rational force—and in view of the further

fact that the myth of AI has the pernicious consequences I also mentioned at the outset, then *why* does it persist?

I think that the persistence of the myth of AI can be explained by a combination of six essentially ideological causal factors: (i) a dogmatic or at least irresponsibly uncritical commitment to the false doctrine of Cartesian dualism, whether substance dualism or property dualism (Hanna and Maiese, 2009), (ii) a dogmatic or at least irresponsibly uncritical commitment to the fantasy of posthumanist or transhumanist spiritualism (see section 2.2), (iii) a dogmatic or at least irresponsibly uncritical commitment to the false doctrine of intellectualism about the nature of rational human animals and rational human cognition (Hanna, 2015a), (iv) a dogmatic or at least irresponsibly uncritical commitment to false (reductive or non-reductive) materialist/physicalist views about the rational human mind, especially including computational functionalism (Hanna and Maiese, 2009), (v) more generally, a dogmatic or at least irresponsibly uncritical commitment to the false mechanistic worldview (Hanna and Paans, 2020; Torday, Miller Jr, and Hanna, 2020; Hanna, 2024), and finally (vi) the hegemony of the military-industrial-digital complex, which amasses and reaps immense wealth and political power precisely by means of effectively and relentlessly disseminating and perpetuating the myth of AI, while at the same time hypocritically issuing public warnings about the dangers of runaway AI:

> Yoshua Bengio, fifty-nine, is the second-most cited living scientist, noted for his foundational work on deep learning. Responding to Page and Sutton, Bengio told me, "What they want, I think it's playing dice with humanity's future. I personally think this should be criminalized." A bit surprised, I asked what exactly he wanted outlawed, and he said efforts to build "AI systems that could overpower us and have their own self-interest by design." In May, Bengio began writing and speaking about how advanced AI systems might go rogue and pose an extinction risk to humanity.
>
> Bengio posits that future, genuinely human-level AI systems could improve their own capabilities, functionally creating a new, more intelligent species. Humanity has driven hundreds of other species extinct, largely by accident. He fears that we could be next—and he isn't alone.
>
> Bengio shared the 2018 Turing Award, computing's Nobel Prize, with fellow deep learning pioneers Yann LeCun and Geoffrey Hinton. Hinton, the most cited living scientist, made waves in May when he resigned from his senior role at Google to more freely sound off about the possibility that future AI systems could wipe out humanity. Hinton and Bengio are the two most prominent AI researchers to join the "x-risk" community. Sometimes referred to as AI safety advocates or doomers, this loose-knit group worries that AI poses an existential risk to humanity.
>
> In the same month that Hinton resigned from Google, hundreds of AI researchers and notable figures signed an open letter stating, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Hin-

> ton and Bengio were the lead signatories, followed by OpenAI CEO Sam Altman and the heads of other top AI labs.
>
> Hinton and Bengio were also the first authors of an October position paper warning about the risk of "an irreversible loss of human control over autonomous AI systems," joined by famous academics like Nobel laureate Daniel Kahneman and Sapiens author Yuval Noah Harari.
>
> LeCun, who runs AI at Meta, agrees that human-level AI is coming but said in a publicdebate against Bengio on AI extinction, "If it's dangerous, we won't build it."
>
> Deep learning powers the most advanced AI systems in the world, from DeepMind's protein-folding model to large language models (LLMs) like OpenAI's ChatGPT. No one really understands how deep learning systems work, but their performance has continued to improve nonetheless. These systems aren't designed to function according to a set of well-understood principles but are instead "trained" to analyze patterns in large datasets, with complex behavior—like language understanding—emerging as a consequence. AI developer Connor Leahy told me, "It's more like we're poking something in a Petri dish" than writing a piece of code. The October position paper warns that "no one currently knows how to reliably align AI behavior with complex values."
>
> In spite of all this uncertainty, AI companies see themselves as being in a race to make these systems as powerful as they can—without a workable plan to understand how the things they're creating actually function, all while cutting corners on safety to win more market share. Artificial general intelligence (AGI) is the holy grail that leading AI labs are explicitly working toward. AGI is often defined as a system that is at least as good as humans at almost any intellectual task. It's also the thing that Bengio and Hinton believe could lead to the end of humanity.
>
> Bizarrely, many of the people actively advancing AI capabilities think there's a significant chance that doing so will ultimately cause the apocalypse. A 2022 survey of machine learning researchers found that nearly half of them thought there was at least a 10 percent chance advanced AI could lead to "human extinction or [a] similarly permanent and severe disempowerment" of humanity. Just months before he cofounded OpenAI, Altman said, "AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies." (Lovely, 2024: pp. 67–68)

For more on this, see sub-sections 2.9.4 and 2.9.5 below.

In any case, the first five factors are classically philosophical factors, so they're at least in principle open to critical rational argumentation, refutation, and correction. But the same isn't the case with respect to the sixth factor, which is essentially social-institutional and political in nature. Sadly, however, I think it's more than merely reasonable to hold that the hegemony of the military-industrial-digital complex is the principal cause of the persistence of the myth of AI, as Garrison Lovely cogently and crisply puts it:

> The debate playing out in the public square may lead you to believe that we have to choose between addressing AI's immediate harms and its inherently speculative existential risks. And there are certainly trade-offs that require careful consideration.

> But when you look at the material forces at play, a different picture emerges: in one corner are trillion-dollar companies trying to make AI models more powerful and profitable; in another, you find civil society groups trying to make AI reflect values that routinely clash with profit maximization.
>
> In short, it's capitalism versus humanity. (Lovely, 2024: p. 79)

Nevertheless, radical social-institutional and political change for the better, or even the best, is in fact really possible, by means of what Michelle Maiese and I have called *the mind-body politic* and *the enactive-transformative principle* (Maiese and Hanna, 2019). So to end this sub-section on an upbeat note, there's at least *some* rational hope for debunking the myth of AI, by devolving and enactively transforming the military-industrial-digital complex.

2.9.4 Oppenheimer, Kaczynski, Shelley, and Frankenscience

> What we are creating now [with the atomic bomb] is a monster whose influence is going to change history, provided there is any history left, yet it would be impossible not to see it through, not only for the military reasons, but it would also be unethical from the point of view of the scientists not to do what they know is feasible, no matter what terrible consequences it may have. (John von Neumann, as reported by his wife and quoted in Dyson, 2012: p. 62.)

What do J. Robert Oppenheimer (1904–1967) and Theodore Kaczynski (1942–2023) have in common?[19] They were both exceptionally intelligent, indeed brilliant, and sometimes troubled, people. They both held BAs from Harvard and received their PhDs very early, at the ages of 23 and 25 respectively. They were both formal scientists: a theoretical physicist and a mathematician respectively. They both taught at UC-Berkeley. Oppenheimer was the father of the atomic bomb, aka the A-bomb, which killed at least 110,000 people at Hiroshima and Nagasaki in 1945, bringing about Japan's unconditional surrender and ending World War II. And Kaczynski was the father of the unabomb (a neologism deriving from an FBI case file labeled "university and airline bomber"), a letter bomb campaign against technocrats that killed three people and injured 23 others between 1978 and 1995.

There are also important dissimilarities. Oppenheimer was never accused of or charged with war crimes, and indeed, until his run-in with McCarthyism in the 1950s, he was universally regarded as a national public hero of science. Kaczyn-

19 For more information about Oppenheimer and Kaczynski, see these two excellent documentaries: Else, 1980; Dammbeck, 2003.

ski, by contrast, was arrested in 1996, sentenced to life imprisonment without parole in 1998, and ultimately committed suicide in prison in June 2023. But most importantly of all, at least during World War II, Oppenheimer fervently affirmed, whereas from the 1970s onwards Kaczynski violently opposed, what I'll call *the thesis of frankenscience*, formulated by another Manhattan Project member, John von Neumann, as per the epigraph at the top of this section. For the purposes of this essay, I'll slightly reformulate the thesis of frankenscience as follows: scientists are always obligated, as scientists, to do whatever they know is feasible, no matter how morally objectionable this is and no matter what bad consequences this may have for humankind.

The neologism "frankenscience" is obviously a riff on the title of Mary Shelley's brilliant and world-famous 1818 novella, *Frankenstein; Or, the Modern Prometheus*, which of course is about a scientist, Victor Frankenstein, who creates a humanoid from corpses and brings it to life with electricity, only to discover that he's created a monster. In turn, Frankenstein's recognition of his tragic mistake can be taken to be Shelley's direct rejection of frankenscience:

> Learn from me, if not by my precepts, at least by my example, how dangerous is the acquirement of [scientific] knowledge, and how much happier that man is who believes his native town to be the world, than he who aspires to become greater than his nature will allow. (Shelley, 1818: vol. 1, ch. 3)

I strongly agree with Shelley. It's crucial to recognize, however, that it's entirely possible to be pro-science, but also anti-frankenscience, and indeed that's my view. Much science is intrinsically valuable, morally unobjectionable, and also has good consequences for humankind. But scientists are never obligated, as scientists, to do whatever they know is feasible, whenever this is morally objectionable or has bad consequences for humankind. On the contrary, they're always morally obligated to refuse to participate in scientific research of these kinds and also to oppose the use of science for such purposes, for the sake of sufficient respect for human dignity. Correspondingly, as I said, I strongly agree with Shelley's direct rejection of frankenscience, whatever her views might have been about science per se. In any case, in this chapter I've been extending that direct rejection of frankenscience to the strong AI research program.

In an important paper released in mid-December 2023, an OpenAI research team—called "The Superalignment Generalization Team"—discussed what they call "the superalignment problem." Here's how OpenAI described it in an overview post on their main website:

> We believe superintelligence—AI vastly smarter than humans—could be developed within the next ten years. However, we still do not know how to reliably steer and control superhuman AI systems. Solving this problem is essential for ensuring that even the most advanced AI systems in the future remain safe and beneficial to humanity….
>
> Current alignment methods, such as reinforcement learning from human feedback (RLHF), rely on human supervision. However, future AI systems will be capable of extremely complex and creative behaviors that will make it hard for humans to reliably supervise them. For example, superhuman models may be able to write millions of lines of novel—and potentially dangerous—computer code that would be very hard even for expert humans to understand.
>
> Relative to superhuman AI models, humans will be "weak supervisors." This is a core challenge for AGI alignment: how can weak supervisors trust and control substantially stronger models? (OpenAI, 2025b)

In other words, the superalignment problem is that Large Language Models (LLMs) like ChatGPT will soon become so massively complex and prodigious that even the people who construct and run them won't be able to comprehend or control them. In turn, The Superalignment Generalization Team's proposed solution to the superalignment problem is what they call "weak-to-strong generalization":

> Our setup
>
> To make progress on this core challenge, we propose an analogy we can empirically study today: can we use a smaller (less capable) model to supervise a larger (more capable) model?

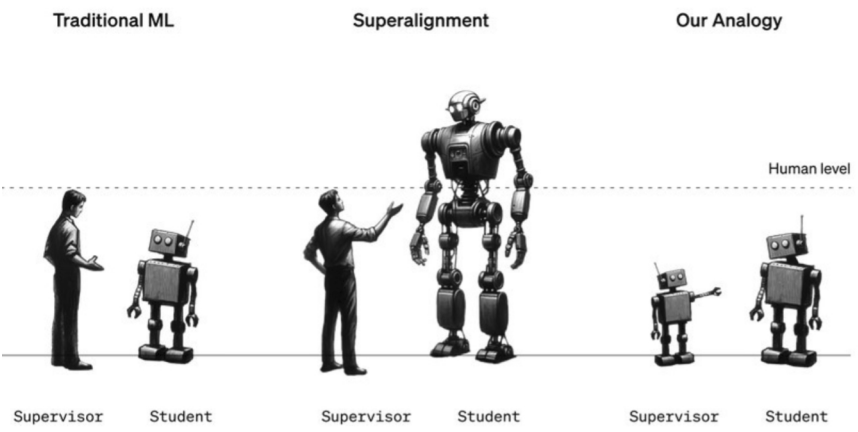
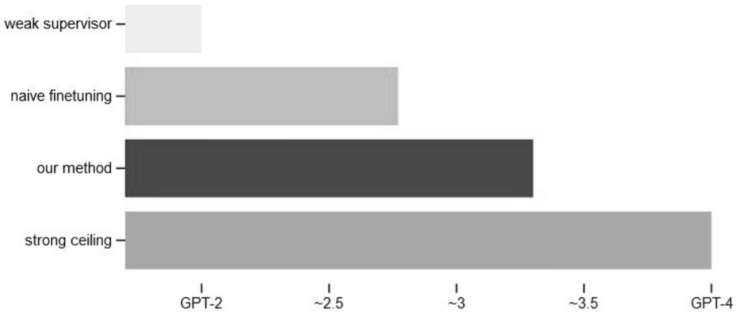


Fig. 2: OpenAI, 2025b.

Naively, we might not expect a strong model to perform better than the weak supervisor that provides its training signal—it may simply learn to imitate all the errors the weak supervisor makes. On the other hand, strong pretrained models have excellent raw capabilities—we don't need to teach them new tasks from scratch, we just need to elicit their latent knowledge. The critical question is then: will the strong model generalize according to the weak supervisor's underlying intent—leveraging its full capabilities to solve the task even on difficult problems where the weak supervisor can only provide incomplete or flawed training labels?

Our results



Typical weak-to-strong generalization across NLP benchmarks: We use a GPT-2-level model as a weak supervisor to finetune GPT-4.

Fig. 3: OpenAI, 2025b.

We can significantly improve generalization in many settings. We use a simple method that encourages the strong model to be more confident—including confidently disagreeing with the weak supervisor if necessary. When we supervise GPT-4 with a GPT-2-level model using this method on NLP tasks, the resulting model typically performs somewhere between GPT-3 and GPT-3.5. We are able to recover much of GPT-4's capabilities with only much weaker supervision.

This method is a proof of concept with important limitations; for example, it still doesn't work on ChatGPT preference data. However, we also find signs of life with other approaches, such as optimal early stopping and bootstrapping from small to intermediate to large models.

Collectively, our results suggest that (1) naive human supervision—such as reinforcement learning from human feedback (RLHF)—could scale poorly to superhuman models without further work, but (2) it is feasible to substantially improve weak-to-strong generalization. (OpenAI, 2025b)

More specifically and technically, however, here's the abstract for The Superalignment Generalizations Team's research paper:

> Widely used alignment techniques, such as reinforcement learning from human feedback (RLHF), rely on the ability of humans to supervise model behavior—for example, to evaluate whether a model faithfully followed instructions or generated safe outputs. However, future superhuman models will behave in complex ways too difficult for humans to reliably evaluate; humans will only be able to weakly supervise superhuman models. We study an analogy to this problem: can weak model supervision elicit the full capabilities of a much stronger model? We test this using a range of pretrained language models in the GPT-4 family on natural language processing (NLP), chess, and reward modeling tasks. We find that when we naively finetune strong pretrained models on labels generated by a weak model, they consistently perform better than their weak supervisors, a phenomenon we call weak-to-strong generalization. However, we are still far from recovering the full capabilities of strong models with naive finetuning alone, suggesting that techniques like RLHF may scale poorly to superhuman models without further work. We find that simple methods can often significantly improve weak-to-strong generalization: for example, when finetuning GPT-4 with a GPT-2-level supervisor and an auxiliary confidence loss, we can recover close to GPT-3.5-level performance on NLP tasks. Our results suggest that it is feasible to make empirical progress today on a fundamental challenge of aligning superhuman models. (Burns et al., 2023: p. 1)

In other words, The Superalignment Generalization Team's proposed solution to the superalignment problem is to use less complex and less prodigious LLMs in order to comprehend and control the more complex and more prodigious LLMs.

But there's an obvious philosophical objection to this methodological strategy. Let's suppose, as I've argued in this section, that it's actually the case that using the currently less complex and less prodigious LLMs that are being aligned by reinforcement learning from human feedback, is already rationally unjustified and immoral. Then how could using those LLMs ever possibly comprehend and control the massively more complex and prodigious future LLMs in a rationally and morally justified way? On the last page of the paper, the team says:

> No need to solve human values. We assume we do not need to solve hard philosophical questions of human values and value aggregation before we can align a superhuman researcher model well enough that it avoids egregiously catastrophic outcomes. (Burns et al., 2023: p. 49)

In other words, implementing The Superalignment Generalization Team's solution will be completely unconstrained by any annoying and distracting philosophical worries to the effect that what OpenAI and other LLM research programs are currently doing is in fact already rationally unjustified and immoral—indeed, as rationally unjustified and immoral as the US government's Manhattan Project during World War II for building, testing, and using the atomic bomb.

2.9.5 Hinton & Me

In an open letter published online in late March 2023, directed not only to the digital technology and AI community in particular but also to the world more generally, card-carrying members of the military-industrial-digital complex—including Elon Musk and Steve Wozniak—urged "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4" (Future of Life Institute, 2023). Here's their letter in full, leaving out the Notes and References:

> AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs. As stated in the widely-endorsed Asilomar AI Principles, *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.* Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one—not even their creators—can understand, predict, or reliably control.
>
> Contemporary AI systems are now becoming human-competitive at general tasks, and we must ask ourselves: *Should* we let machines flood our information channels with propaganda and untruth? *Should* we automate away all the jobs, including the fulfilling ones? *Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable. This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's recent statement regarding artificial general intelligence, states that *"At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models."* We agree. That point is now.
>
> Therefore, we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.
>
> AI labs and independent experts should use this pause to jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts. These protocols should ensure that systems adhering to them are safe beyond a reasonable doubt. This does *not* mean a pause on AI development in general, merely a stepping back from the dangerous race to ever-larger unpredictable black-box models with emergent capabilities.
>
> AI research and development should be refocused on making today's powerful, state-of-the-art systems more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal.

> In parallel, AI developers must work with policymakers to dramatically accelerate development of robust AI governance systems. These should at a minimum include: new and capable regulatory authorities dedicated to AI; oversight and tracking of highly capable AI systems and large pools of computational capability; provenance and watermarking systems to help distinguish real from synthetic and to track model leaks; a robust auditing and certification ecosystem; liability for AI-caused harm; robust public funding for technical AI safety research; and well-resourced institutions for coping with the dramatic economic and political disruptions (especially to democracy) that AI will cause.
>
> Humanity can enjoy a flourishing future with AI. Having succeeded in creating powerful AI systems, we can now enjoy an "AI summer" in which we reap the rewards, engineer these systems for the clear benefit of all, and give society a chance to adapt. Society has hit pause on other technologies with potentially catastrophic effects on society. We can do so here. Let's enjoy a long AI summer, not rush unprepared into a fall. (Future of Life Institute, 2023)

I strongly disagree with what's being asserted and argued in this letter. For people with a huge vested interest in designing, producing, and marketing digital technology and AI, to call on AI labs and the fabulously wealthy global technocratic capitalist corporations (including the corporations they own or founded, for example, Musk/Tesla and Wozniak/Apple) that design, produce, and market digital technology, to "pause" their research on Large Language Models (LLMs) and the chatbots that implement these models "beyond GPT-4" for "at least six months,"[20] *and then* go forward again full-steam-ahead, having paid public lip-service to "digital ethics," so that these corporations can create an all-out existential attack on humankind, is just like it would have been for leading US generals during the latter phases of World War II to call on the Manhattan Project to "pause" their work on the atomic bomb for "at least six months," then go forward again full-steam-ahead, having paid public lip-service to "just war theory," so that the USA could drop the A-bomb not just once but twice on hundreds of thousands of Japanese civilians, and thereby create a permanent existential threat for humankind.

On 1 May 2023—significantly, May Day—Geoffrey Hinton, a groundbreaking researcher on neural networks and LLMs, quit his job at Google and then publicly

20 The authors and signatories of the open letter say that "[i]f such a [six-months-minimum] pause cannot be enacted quickly, governments should step in and institute a moratorium" (Future of Life Institute, 2023). Now, "moratorium" means "temporary prohibition of an activity." So, supposing that a self-adopted pause or government moratorium were to occur, then this still logically or at least conversationally implies that sooner or later, giant AI experiments and LLM/chatbot technology will be resumed. On the contrary, I'm saying that it should be shut down permanently in its infancy, just as A-bomb research and nuclear weapons technology should have been shut down permanently in its infancy.

made a very similar point, also explicitly using the analogy with the Manhattan Project:

> As companies improve their A.I. systems, [Hinton] believes, they become increasingly dangerous. "Look at how it was five years ago and how it is now," he said of A.I. technology. "Take the difference and propagate it forwards. That's scary."
>
> Until last year, he said, Google acted as a "proper steward" for the technology, careful not to release something that might cause harm. But now that Microsoft has augmented its Bing search engine with a chatbot—challenging Google's core business—Google is racing to deploy the same kind of technology. The tech giants are locked in a competition that might be impossible to stop, Dr. Hinton said.
>
> His immediate concern is that the internet will be flooded with false photos, videos and text, and the average person will "not be able to know what is true anymore."
>
> He is also worried that A.I. technologies will in time upend the job market. Today, chatbots like ChatGPT tend to complement human workers, but they could replace paralegals, personal assistants, translators and others who handle rote tasks. "It takes away the drudge work," he said. "It might take away more than that."
>
> Down the road, he is worried that future versions of the technology pose a threat to humanity because they often learn unexpected behavior from the vast amounts of data they analyze. This becomes an issue, he said, as individuals and companies allow A.I. systems not only to generate their own computer code but actually run that code on their own. And he fears a day when truly autonomous weapons—those killer robots—become reality.
>
> "The idea that this stuff could actually get smarter than people—a few people believed that," he said. "But most people thought it was way off. And I thought it was way off. I thought it was 30 to 50 years or even longer away. Obviously, I no longer think that."
>
> Many other experts, including many of his students and colleagues, say this threat is hypothetical. But Dr. Hinton believes that the race between Google and Microsoft and others will escalate into a global race that will not stop without some sort of global regulation.
>
> But that may be impossible, he said. Unlike with nuclear weapons, he said, there is no way of knowing whether companies or countries are working on the technology in secret. The best hope is for the world's leading scientists to collaborate on ways of controlling the technology. "I don't think they should scale this up more until they have understood whether they can control it," he said.
>
> Dr. Hinton said that when people used to ask him how he could work on technology that was potentially dangerous, he would paraphrase Robert Oppenheimer, who led the U.S. effort to build the atomic bomb: "When you see something that is technically sweet, you go ahead and do it."
>
> He does not say that anymore. (Metz, 2023)

What Hinton (apparently) and I (absolutely) strongly believe, then, is that we ought to ban all giant AI experiments and LLM/chatbot technology while they are still in

their infancy, just as we ought to have banned all atomic bomb experiments and nuclear weapons technology while they were still in their infancy. My strong belief, in turn, is a direct expression and logical implication of the doctrine I call *dignitarian neo-Luddism with respect to digital technology*, which I'll spell out and justify in chapters 6 and 7 below.

For the time being, however, what I'm asserting is that we all ought to be dignitarian neo-Luddites with respect to digital technology. Why? To be sure, there are many ways in which digital technology can be bad and wrong in the dignitarian sense, including invasive digital surveillance, digitally-driven weapons and weapon systems, algorithmic bias, and digital manipulation and nudging. And of course there are also ways in which digital technology can be bad and wrong in the utilitarian sense, for example, by putting many people out of work, and by substantially contributing to climate change disaster. But the principal reason for being a dignitarian neo-Luddite with respect to digital technology is that our excessive use of and indeed addiction to digital technology is systematically undermining our innate capacities for thinking, caring, and acting for ourselves. This is preeminently true with respect to the new chatbots—for example, ChatGPT and LaMDA—and the myth of AI more generally, but also to an increasingly important degree true for our excessive use of and addiction to smart-phones, desktop and laptop computers, the internet, social media, and so on and so forth. When you combine our excessive use of and addiction to chatbots and AI with our excessive use of and addiction to smart-phones, desktop and laptop computers, the internet, social media, etc., the result is nothing less than an all-out existential attack on our rational human mindedness or intelligence. The unified set of innate capacities that constitutes our rational human mindedness or intelligence in turn constitutes our human real personhood, which again in turn is the metaphysical ground of our human dignity (see chapters 3–5). Therefore, this all-out existential attack on our rational human mindedness or intelligence is also an all-out existential attack on our human dignity.

The Cassandra-like prophecy and warning that I'm issuing, however, is not that chatbots or AI more generally could ever become rational, super-intelligent, and morally satanic, and then run amok. Interestingly, that seems to be the main source of Bengio's and Hinton's worries. But in fact, it's strongly metaphysically or synthetic a priori impossible for computing machines ever to be rationally minded or intelligent in the sense that we're rationally minded or intelligent, because (i) it's strongly metaphysically or synthetic a priori necessary that all creatures possessing the seven basic innate capacities I listed in section 2.7 above are complex living organisms, i. e., animals (Hanna and Maiese, 2009), hence not machines, hence not computing machines, and (ii) it's also strongly metaphysically or synthetic a priori necessary that our rational mindedness or intelligence in-

cludes (iia) an innate non-basic ("non-basic," in the sense that it essentially depends on the seven basic innate capacities listed in the just-previous paragraph) capacity for spontaneous creativity, and also (iib) an innate non-basic capacity for either conceptual or essentially non-conceptual a priori intuition of (iib1) innately-specified universal, unconditional, a priori or non-empirical moral principles such as everyone ought always to choose and act with sufficient respect for everyone's dignity, including their own, (iib2) universal, unconditional, a priori or non-empirical logical principles such as the minimal principle of non-contradiction, namely, not every statement is both true and false, and (iib3) the universal a priori or non-empirical formal structures of the orientable, three-dimensional space and the forward-directed, processual, purposive, asymmetric organic time in which our minded animal bodies are ineluctably embedded (Hanna, 2006a, 2015a, 2018a, 2018b), none of which can ever exist in computing machinery. And, closely related to these strongly metaphysically or synthetic a priori necessary modal facts, there are also various other linguistic, logical, mathematical, and metaphysical reasons why computing machinery can never be rationally minded or intelligent in the sense that we're rationally minded or intelligent, as per the nine arguments I presented in section 2.9.2 (see also Chomsky, Roberts, and Watamull, 2023; Keller, 2023; Landgrebe and Smith, 2022).

On the contrary, the Cassandra-like prophecy and warning that I'm issuing about this all-out existential attack on our rational human mindedness or intelligence is instead directed at the global technocratic capitalist corporations—especially those that supply weapons and surveillance systems for military and government use—millionaires, and billionaires who reap immense profits and wield immense political power by designing, producing, marketing, and above all controlling our use of and reliance on digital technology: namely, the members of the military-industrial-digital complex. Correspondingly, my Cassandra-like prophecy and warning is simply this:

> The members of the military-industrial-digital complex are systematically harming and oppressing ordinary people like us by not only enabling but also effectively mandating our excessive use of and addiction to digital technology, which in turn systematically undermines our innate capacities for thinking, caring, and acting for ourselves, and therefore undermines our human real personhood, and thereby violates our human dignity—therefore, we ought to ban all giant AI experiments and LLM/chatbot technology while they are still in their infancy, just as we ought to have banned all atomic bomb experiments and nuclear weapons technology while they were still in their infancy.

2.10 Organoid Intelligence? Just Say *No*

Just when, as per sections 2.2 to 2.9, you thought (i) that you were beginning to wrap your head around Turing's transhumanist spiritualist fantasies, the new LLMs or chatbots, The Turing Test, Robobob, Bob, the falsity and impossibility of strong AI, and the myth of AI, and therefore also thought (ii) that you could sleep comfortably again at night, perhaps dreaming of electric sheep (Dick, 1968), then you'll no doubt be dismayed to learn that contemporary scientists working on the foundations of digital technology have yet another surprise up their sleeves. For a new, well-funded, and mind-blowingly—or brain-blowingly—revolutionary multidisciplinary research program in biological computing called Organoid Intelligence, aka OI, was published in *Frontiers in Science* in late February 2023:

> Recent advances in human stem cell-derived brain organoids promise to replicate critical molecular and cellular aspects of learning and memory and possibly aspects of cognition in vitro. Coining the term "organoid intelligence" (OI) to encompass these developments, we present a collaborative program to implement the vision of a multidisciplinary field of OI. This aims to establish OI as a form of genuine biological computing that harnesses brain organoids using scientific and bioengineering advances in an ethically responsible manner. Standardized, 3D, myelinated brain organoids can now be produced with high cell density and enriched levels of glial cells and gene expression critical for learning. Integrated microfluidic perfusion systems can support scalable and durable culturing, and spatiotemporal chemical signaling. Novel 3D microelectrode arrays permit high-resolution spatiotemporal electrophysiological signaling and recording to explore the capacity of brain organoids to recapitulate the molecular mechanisms of learning and memory formation and, ultimately, their computational potential. Technologies that could enable novel biocomputing models via stimulus-response training and organoid-computer interfaces are in development. We envisage complex, networked interfaces whereby brain organoids are connected with real-world sensors and output devices, and ultimately with each other and with sensory organ organoids (e.g. retinal organoids), and are trained using biofeedback, big-data warehousing, and machine learning methods….The many possible applications of this research urge the strategic development of OI as a scientific discipline. We anticipate OI-based biocomputing systems to allow faster decisionmaking, continuous learning during tasks, and greater energy and data efficiency. Furthermore, the development of *"intelligence-in-a-dish"* could help elucidate the pathophysiology of devastating developmental and degenerative diseases (such as dementia), potentially aiding the identification of novel therapeutic approaches to address major global unmet needs. (Smirnova et al., 2023: p. 2, italics added)

The shudder-quoted neologism, "intelligence-in-a-dish," will remind philosophers and other philosophically-minded people of Hilary Putnam's well-known "Brain-in-a-Vat," aka BIV, thought-experiment in *Reason, Truth, and History* (Putnam, 1981: pp. 1–21). And indeed, the powerful thought-shaping image (Hanna and Paans, 2021) of a living human brain floating listlessly in a dish or vat yet also fully interfaced with sophisticated digital technology, is essentially the same for

OI and BIV. But there's an important and indeed categorical difference between the two scenarios. According to Putnam's BIV scenario, we're asked to imagine the skeptical possibility that unbeknownst to us, we as we currently are, as rational human minded animals, i. e., human real persons, might be nothing but brains in a vat, and also, like the fictional scenario presented 18 years later in the 1999 blockbuster dystopian science fiction movie directed by Lana and Lilly Wachowski, The Matrix, that nothing really is what it seems to us to be, precisely because it's nothing but, as we would say nowadays, a "virtual reality," or a computer-generated living dream created by evil computer scientists, the 20th-century equivalents of Descartes's evil demon. OI is also superficially similar to, yet categorically distinct from, the fictional scenario presented in Ridley Scott's brilliant 1982 dystopian science fiction film, *Blade Runner*—itself based on Philip K. Dick's equally brilliant 1968 dystopian science fiction novel, *Do Androids Dream of Electric Sheep?* (Dick, 1968)—describing the "Nexus VI replicants," artificially manufactured and marketed by the "Tyrell Corporation," such that this particular generation of Nexus replicants are relatively short-lived biophysical counterparts of human real persons, i. e., bio-engineered, synthetic humanoid real persons, possessing all of the innate capacities constituting our rational human mindedness, including consciousness and self-consciousness, but used for industrial and other technocratic global corporate capitalist purposes, treated as slaves, and permissibly "retired" at will by hired assassins called "blade runners."

To summarize, BIVs are identical to us, supposedly metaphysically reduced to our brains, and filled to the brim with false beliefs about the world and ourselves; and Nexus VI replicants, although not precisely identical to us, i. e., rational minded members of the species *homo sapiens*, are nevertheless people just like us, who are also being enslaved and oppressed by us, merely because they've been produced as fully-grown in factories, but not, like us, created as fertilized eggs, then gestated as fetuses, born as neonates from wombs, and then nurtured through all developmental phases from infancy to adulthood, in the world. But in categorical contradistiction to BIVs and Nexus VI replicants alike, OI is the scientific research program scenario for growing new brain organoids from human brain cells, then synchronously embedding those cultured organoids inside cutting edge digital technology, including chatbots and robots, and then, using machine-learning programs, training up those computational systems to the level of contemporary chatbots or beyond. As the scientists developing the OI research program put it:

> We envision using biofeedback to systematically train organoids with increasingly complex sensory inputs and output opportunities—interfacing the brain organoids with computers, sensors, and machine interfaces to facilitate supervised and unsupervised learning. We use

> the term "OI" for this approach to stress its complementarity to AI—where computers aim to perform tasks done by brains, often by modeling our understanding of learning. However, *while AI aims to make computers more brain-like, OI research will explore how a 3D brain cell culture can be made more computer-like.* (Smirnova et al., 2023: p. 4, italics added)

Now, from a broadly Kantian metaphysical and moral point of view, what should we say about OI?

Here are my basic metaphysical commitments. By *rational human mindedness* or *human intelligence* I mean the essentially embodied (i. e., necessarily and completely embodied in a suitably complex living animal organism), unified set of basic innate cognitive, affective, and practical capacities that are present in all and only those human animals possessing the essentially embodied neurobiological basis of those capacities, namely: consciousness, self-consciousness, caring, sensible cognition, intellectual cognition, volition, and free agency (Hanna and Maiese, 2009; Hanna, 2011a, 2015a, 2018b—see also section 2.7 above). This unified set of cognitive, affective, and practical capacities constitutes our human real personhood, which in turn is the metaphysical ground of our human dignity (Hanna, 2018b, 2018c; see also chapter 3). Metaphysically speaking, then, I'm committed to the view that rational human mindedness or human intelligence must be alive and organismic; hence since no machine is alive and organismic, and since all computers are essentially Turing machines, then no computer can be rationally minded or genuinely intelligent in the sense in which we're intelligent. I'm also committed to the view that there are several strong purely logical and mathematical, topological, and psycholinguistic reasons why no computer, no matter how sophisticated, can ever be intelligent and think in the sense that we are intelligent and think (Hanna, 2006a; sections 2.1 – 2.8 above; Keller, 2023; Landgrebe and Smith, 2022). And I'm also committed to the view that rational mindedness, or genuine intelligence, is necessarily and completely embodied in suitably complex living animal organisms, whether human or non-human.

As we've seen, the OI research program seeks to create computational systems that include embedded, synchronous organoids grown from human brain cells. But in view of my metaphysical commitments, the following three claims seem self-evidently true to me. First, a digital computer that merely includes some organic parts is still a machine, hence it cannot be alive, and therefore no actual or possible OI can ever be rationally minded or genuinely intelligent in the sense in which we're rationally minded and intelligent. Second, logically and mathematically speaking, a computer is still nothing but a computer, even if it contains some organic parts: hence no actual or possible OI can ever be rationally minded or genuinely intelligent in the way we are. And third, since the organic parts of an OI are only parts of, or even the whole of, a human brain, and do not constitute a

complete living rational minded or genuinely intelligent animal, whether human or non-human, then no actual or possible OI is ever completely embodied, therefore no actual or possible OI has rational mindedness or genuine intelligence like ours. In short, just as every actual or possible AI is artificial, but not intelligent, so too every actual or possible OI will be organoid, but not intelligent.

What about the moral implications of OI in relation to my basic moral commitments? Here's what the scientists say:

> [M]oral attitudes toward OI may depend less on epistemological concerns mentioned above, such as the role of specific cognitive capacities in assessments of moral status, and more on ontological arguments of what constitutes a human being. Perceptions of (re)creating 'human-like' entities in the lab are likely to evoke concerns about infringing on human dignity that could reflect secular or theological beliefs about the 'essential' nature of the human being. (Smirnova et al., 2023: p. 15)

As I pointed out in chapter 1, the core thesis of Dignitarian Digital Ethics, aka DDE, is Dignitarian Cybernetic Instrumentalism, aka DCI, which says that digital technology, no matter how powerful, sophisticated, or profitable, is nothing more and nothing less than a tool created by humankind for the betterment of humankind, whose use therefore not only can but also should be strictly constrained by general and specific moral principles flowing from the concept and fact of human dignity. Since, as we'll see in section 3.4 below, my account of the metaphysics of human dignity explicitly grounds dignity on human real personhood, which in turn is constituted by "the role[s] of specific cognitive capacities in assessments of moral status" (Smirnova et al., 2023: p. 15), then my account does not depend on "ontological arguments of what constitutes a human being," such that

> [p]erceptions of (re)creating 'human-like' entities in the lab are likely to evoke concerns about infringing on human dignity that could reflect secular or theological beliefs about the 'essential' nature of the human being. (Smirnova et al., 2023: p. 15)

Therefore, since according to my account, no OI can ever, even in principle, be a human real person with dignity, then necessarily, the OI research program, in and of itself, will never involve "infringing on human dignity" (Smirnova et al., 2023: p. 15). Of course, if the OI research team used coercive means to obtain brain cells from some human real persons—say, at the point of a gun, or by means of an immoral social-institutional system for trafficking in human brain cells—then that would violate human dignity; but it wouldn't flow from the OI research program per se.

Nevertheless, just like LaMDA, Bard, Sydney, ChatGPT, and all the other chatbots to-be, unless we universally enact DCI, OI will enable and indeed mandate our

excessive use of and indeed addiction to it. In turn, this will systematically undermine our innate capacities for thinking, caring, and acting for ourselves. Then, when you combine our excessive use of and addiction to chatbots and OI with our excessive use of and addiction to smart-phones, desktop and laptop computers, the internet, social media, and so on, the result is nothing less than what I'll call *the invasion of the mind snatchers*, riffing on the title of Don Siegel's famous 1956 science fiction/horror film, *Invasion of the Body Snatchers.* More precisely even though OI will never be intelligent in the sense in which we're intelligent, nor will it ever have human dignity, not even in principle, nevertheless, just like the mind-snatching chatbots, unless we universally enact DCI, OI will be a serious threat to the well-being of humankind by oppressing us and indeed outright violating our human dignity. One of the direct implications of DCI is that if any line of research in any of the formal or natural sciences that theoretically feed digital technology, is implemented in a form of digital technology that either fails to respect human dignity sufficiently or outright violates human dignity, then we should refuse to pursue that line of research, no matter how massively power-enhancing or massively profit-making this form of digital technology will be for those who belong to what I've called "the military-industrial-digital complex" (see chapter 1, esp. n. 2, above). Therefore, from the standpoint of DCI, it's a moral categorical imperative that we not pursue the OI research program.

Organoid Intelligence? Metaphysically and morally, just say *no.* I'll come back again to this general line of critical argument in chapters 6 and 7.

3 What's So Special About Human Dignity? Its Metaphysics

3.1 Introduction

What's so special about human dignity (Etinson, 2020)? Well, to focus on our recent and contemporary world for a moment, perhaps you're wondering whether, and if so why, you should have been wearing a mask, social-distancing, isolating or sheltering, and taking other reasonable precautions during the early and middle stages of the COVID-19 pandemic, even though it was annoying, depressing, inconvenient, stale, flat, and/or unprofitable to do these things, and even though, at least in the middle stages and beyond, you and many others had been fully vaccinated and perhaps also boostered, worldwide. Well, *yes*, you should have been doing these things; and the principal reason for doing them then, and again in the future, if another pandemic should arise—leaving aside either narrow self-interest or publicly-oriented utilitarian calculations—is that it's morally required by sufficient respect for everyone's human dignity. It's morally wrong to fail to do those things and thereby to put other people and/or yourself at significant risk of contracting and spreading the deadly disease and its variants, principally because you're thereby violating their human dignity and yours.

But, philosophically backpedaling for a moment, does human dignity really exist? If so, then what is its nature and how is that nature grounded, what are its essential moral implications, how do we know them, and how can this dignitarian knowledge be applied in real-world political contexts? No questions could be more important for humankind. Therefore, it's a rational human imperative to provide a clear, distinct, consistent, complete, and true—or at least philosophically intelligible, defensible and plausible—theory of human dignity.

By way of a preview, here are capsulized versions of the answers I'll give to those questions in the rest of this chapter and chapter 4 below.

First, human dignity really exists because (i) no one, not even a dignity-skeptic, could give their actual or possible rational consent to being treated either as a mere means to someone's ends (i. e., their desires or goals) or as a mere thing, and (ii) its being absolutely impermissible to treat any people (including oneself) either as a mere means or as a mere thing is an essential property of human dignity.

Second, human dignity is the absolute, non-denumerably infinite, intrinsic, and objective value or worth of human real persons as ends-in-themselves, and human real personhood is metaphysically grounded in an essentially embodied,

https://doi.org/10.1515/9783111260525-006

Fig. 4: "*Prüfung/Test*," by Edith Breckwoldt (Hamburg DE, 2004) (Author's photograph, 2019).

unified set of seven innate cognitive, emotional, and practical capacities present in all and only those human animals possessing the essentially embodied neurobiological basis of those capacities (see section 2.7 above).

Third, the essential moral implications of human dignity are an hierarchically-ordered set of (either absolutely or *ceteris paribus*, i. e., obtaining only "other things being equal") universal moral principles specifying ways of always treating all human real persons with sufficient respect for their human dignity, the essence of which is the absolutely universal obligation never to treat any human real person (including oneself) as a mere means or as a mere thing, in a thoroughly non-ideal natural and social world.

Fourth, human dignity and its essential moral implications are known by a multifaceted systematic method that includes (i) essentially reliable a priori moral intuitions of basic principles supplemented by logical rationality and reasoning, (ii) fairly reliable cognitive and practical constructive knowledge of non-basic principles under those basic principles, (iii) considered moral judgments in real-world contexts and in thought-experiments by way of applying and further specifying those basic and non-basic principles, and (iv) empathetic intersubjective moral phenomenology.

Fifth and finally, this dignitarian knowledge can be applied in real-world sociopolitical contexts only by enacting human dignity: that is, only by means of designing, creating, and sustaining all and only specifically dignitarian social institutions that Michelle Maiese and I have called *constructive, enabling* social institutions (Maiese and Hanna, 2019: esp. chs. 1–3 and 6–7).

It should be already obvious that the general theory of human dignity I'm presenting and defending in this and the next chapter, and throughout the rest of this book, in the specific guise of DDE, is essentially Kantian in philosophical inspiration. But although I'll sometimes refer to Kant's writings, this general theory of human dignity is neither intended to be an interpretation of Kant's writings, nor in any way restricted by the requirement to remain consistent with or defend any of Kant's own doctrines (for example, his alleged noumenal realism, hatred of emotions, moral formalism and rigorism, political liberalism, etc.) or his personal prejudices (for example, his alleged racism, sexism, xenophobia, etc.). Therefore—to play a riff on Josiah Royce's pithy definition of idealism: "the world and the heavens, and the stars are all *real*, but not so *damned* real" (Royce, 1970: p. 217)—this general theory of human dignity and its sub-species DDE are both essentially *Kantian* in philosophical inspiration, but not so *damned* Kantian. In other words, this general theory of human dignity and its sub-species DDE involve a creative use of some Kantian ideas that are also independently defensible, and they diverge from either Kant's own writings or orthodox Kantianism whenever that's required by attentiveness to manifest reality and/or critical reflection. In view of the social-

institutional facts I've called *the Kant wars*, one element of which is a widespread anti-Kantian bias in contemporary philosophy (Hanna, 2020a), it's (unfortunately) necessary to make this point explicitly. And in order to emphasize that point, I'll always call these creatively-used Kantian ideas *the broadly Kantian theory of human dignity* and *DDE.*

3.2 How to Refute the Dignity-Skeptic

What I'll call *the dignity-skeptic* is anyone who, for any reason whatsoever, denies the real existence of human dignity. But if one asked the dignity-skeptic the following question,

> "Would you or could you rationally consent to being summarily beheaded merely because of your skin pigmentation, ethnicity, gender, sexual orientation, ability or disability, religious beliefs, or political beliefs?, yes or no, please give reasons for your answer, and please reply in all sincerity,"

and then if the dignity-skeptic complied with those requests, they would obviously answer "of course not!" But any reason sincerely provided by the dignity-skeptic that's sufficient to explain why not, would be necessarily equivalent to appealing to the skeptic's own human dignity, since it would necessarily involve the absolute impermissibility of themselves being treated by someone as a mere means or a mere thing for any purportedly sufficient reason that singles them out only as a mere token of some identity-type, thereby reducing themselves to being a mere means or mere thing under that type. And its being absolutely impermissible to treat any people (including oneself) either as a mere means or as a mere thing is an essential property of human dignity. Therefore, the dignity-skeptic's answer would entail or presuppose the real existence of human dignity in at least themselves. So the dignity-skeptic's answer is self-refuting.

3.3 What Is Human Dignity?

According to the broadly Kantian theory of human dignity, human dignity is the absolute, non-denumerably infinite, intrinsic, and objective value or worth of human real persons as ends-in-themselves, and human real personhood is grounded in a unified set of seven innate cognitive, emotional, and practical capacities present in all and only human animals possessing the essentially embodied neurobiological basis of those capacities (see section 2.7 above). These seven capacities

jointly constitute what I've called *rational human mindedness.* Some human animals are born permanently lacking the essentially embodied neurobiological basis of rational human mindedness or have suffered its permanent destruction by accident, disease, or violent mishap, and therefore some human animals do not have human dignity because they are not human real persons. So not necessarily all human animals are real persons. Conversely, not necessarily all real persons are human: it's really possible for there to be real persons belonging to other animal species, whether on the earth or other planets. If so, then they'll have dignity too.

Nevertheless, and in any case, you're a real person, and so am I. And so is every other living organism that's capable of fully understanding those words, feeling their normative force, and then choosing and acting under the guidance of that normative force. Neither logically possible or conceivable non-animal persons, disembodied persons, or divine persons, nor actual artificial persons (personae) or actual collective persons, created by human convention, are real persons in this sense. For human real persons, like all real persons, are essentially embodied minds (Hanna and Maiese, 2009; Hanna, 2011a). In turn, the essential embodiment thesis has two logically distinct parts: (i) the necessary embodiment of conscious minds like ours in a living organism (the necessity thesis), and (ii) the complete neurobiological embodiment of conscious minds like ours in all the vital systems, vital organs, and vital processes of our living bodies (the completeness thesis). The necessity thesis says that necessarily, conscious minds like ours are alive. Negatively formulated, it says that conscious minds like ours cannot be dead, disembodied, or machines. By contrast, the completeness thesis says that conscious minds like ours are fully spread out into our living organismic bodies, necessarily including the brain, but also necessarily not restricted to the brain. In view of the essential embodiment thesis, specifically human real persons are real persons who are necessarily and completely, human animals, hence we're "human, all too human."

Later in this chapter, I'll work out an explicit metaphysical definition of human real persons (see also Hanna, 2018b: chs. 6–7). Right now, I want to concentrate on the metaphysics of the dignity of human real persons as such, or in Kant's terminology, on *Würde.* To say that human real persons have dignity is to say that they're absolutely, non-denumerably infinitely, intrinsically, and objectively valuable or worthy ends-in-themselves. What, more precisely, do I mean by saying that? Objective values are whatever anyone can care about, that is, whatever anyone can aim her emotions (i.e., desires, feelings, or passions) at. Otherwise put, objective values are what Kant called "ends" (*Zwecke*). In turn, "absolute" means "unconditionally necessary." So to say that human real persons are absolutely, non-denumerably infinitely, intrinsically, objectively valuable or worthy ends-in-themselves, or that they have dignity, is to say that their value as ends-in-themselves is not only

an unconditionally necessary, internal feature of the kind of manifestly real being they are, but also the very highest kind of value or worth.

Why?—Because it's only by means of the unified set of capacities jointly constituting us as human real persons that any sort of consciousness or subjective experience of any sort of value or worth, any sort of recognition of any sort of value or worth, any sort of self-consciousness of one's own value or worth, any sort of free action for the sake of any sort of value or worthy goal, and any sort of creative production of any sort of value or worth, can ever occur. These innate capacities are natural, and human persons are natural, therefore human dignity is natural: no appeal to God or anything non-natural is required. At the same time, however, human dignity is a transfinite value or worth, just as the set of real numbers has transfinite cardinality. So dignity according to this conception is no more "weird" or "mystical" or "unscientific" than Zermelo's and Fraenkel's well-ordered set theory plus the Axiom of Choice,[21] aka ZFC (Zermelo, 1930, 1967a, 1967b, 1967c), is "weird" or "mystical" or "unscientific." Just as ZFC is the foundation of all modern mathematics, so too human dignity is the foundation of all human morality.

Now many things are intrinsically objectively valuable or worthy, or ends-in-themselves—for example, pleasant bodily or sensory experiences, vivid emotional experiences, beautiful natural objects and environments, fine craftsmanship, skillfully-played sports, good science, good philosophy, good works of art, and any job well done. To say that human real persons are absolutely, non-denumerably infinitely, intrinsically, objectively valuable or worthy ends-in-themselves—i. e., that they have dignity—however, is to say that each of us has a moral value or worth that is a transfinite cardinal quantity in relation to all denumerable or countable, economic, or otherwise instrumental kinds of value, for example psychological pleasure or preference-satisfaction. It seems clear that however we measure such things, whether in terms of market value or monetary price, degrees of psychological pleasure, degrees of preference-satisfaction, or comparative rankings of such things, nevertheless every actual or possible economic or otherwise instrumental value or worth is expressible as some rational number quantity or another, including denumerably infinite rational number quantities. Then, by essentially the same method that Georg Cantor used to show the existence of transfinite numbers (Cantor, 1891, 2019) at least in principle, we can create a vertical and denumerably infinite list of every actual or possible economic or otherwise instrumental value or worth, then draw a diagonal across it, and discover another value or worth that's categorically higher than any economic or otherwise instrumental

21 The Axiom of Choice is equivalent to licensing the power-set operation (i. e., constructing the set of all subsets) on finite or infinite sets.

value or worth. So this value or worth is the prime example of what—following Cantor's alternative term for transfinite numbers, *transcendental numbers*—I've called *transcendental normativity* (Hanna, 2015b). Correspondingly, it's what I'll call *transcendental value*, by which I mean either a single transcendental value or else a unified system of several distinct but essentially complementary or interlocking transcendental values. Kant called the unified system of all transcendental values "the highest good." The dignity of human real persons has transcendental value in that sense: thus each human real person, by virtue of their dignity, has transcendental value, and their human dignity also inherently belongs to the unified system of all transcendental values, aka the highest good.

Even though each individual human real person has transcendental value—a transfinite value or worth that's categorically higher than any economic or otherwise instrumental value or worth and thereby irreducible to any economic or otherwise instrumental value—nevertheless the value or worth of groups of human real persons can still be calculated from the cardinality of the membership of the group, just as there is an arithmetic of transfinite cardinal numbers. The dignitarian transcendental value of $N > 1$ human real persons is N times greater than the dignitarian transcendental value of one human real person. That's why it's twice as good to save the lives of two human real persons as it is to save the life of one human real person, and twice as bad for two human real persons to die as it is for one human real person to die, and so on. This special kind of transfinite/transcendental value-calculability is true of groups of human real persons, and *yet* the dignity of each individual human real person has a value or worth that's categorically higher than any economic or instrumental value or worth and thereby irreducible to any economic or otherwise instrumental value or worth, which is to say that it's a transcendental value, and that it belongs to the highest good. The highest good in this sense is in each and every one of us; and that's the one and only sense in which we're all morally and politically equal. But apart from that broadly Kantian dignitarian sense of equality, as Harry Frankfurt has compellingly argued (Frankfurt, 2015), egalitarianism more generally is a misguided and mistaken moral and political ideal. What's strictly and universally morally and politically obligatory is to treat every human real person with sufficient respect for their human dignity,[22] which will sometimes necessarily involve strict equality of treatment across sets of individual human real persons, but not always or necessarily.

Treating a human real person with sufficient respect for their human dignity, in turn, has three individually necessary, individually insufficient, and jointly suf-

22 This is a broadly Kantian dignitarian generalization of what Frankfurt calls "the doctrine of sufficiency": see Frankfurt, 2015: p. 7.

ficient conditions: (i) a human real person is treated with sufficient respect only if they are *not* treated either as a mere means or as mere thing, for example, in the way that Nazis treated people, like a piece of garbage or offal, (ii) a human real person is treated with sufficient respect only if they are treated in such a way that they can give their explicit and/or implicit rational consent[23] to that treatment, and (iii) a human real person is treated with sufficient respect only if they are treated *with kindness*—that is, with benevolent attention to their *true* human needs (Maiese and Hanna, 2019: ch. 3; see also section 5.3 below). These are mutually logically distinct and individually necessary, but still *individually insufficient* conditions for sufficient respect for human dignity. For, despite initial appearances, they're not necessarily equivalent, for two reasons.

First, it's at least minimally really possible for a human real person to give their explicit or implicit rational consent to being treated either as a mere means or as a mere thing. Indeed, it's at least minimally really possible that a human real person could explicitly or implicitly rationally consent even to becoming someone else's slave or to being killed by that other person—as an extreme form of self-abasement, self-punishment, self-sacrifice, or sexual self-expression. One real-world example is the notorious "German cannibals" case in 2002 (BBC News, 2002). I do hold, however, that the German cannibals' explicit rational consent also violates their implicit rational consent. But the more general point I am making here is that in all such cases, someone, of their own free will, disrespects themselves and therefore is choosing and acting impermissibly.

Second, even if a human real person is not being treated as a mere means or as a mere thing, and can also give their explicit or implicit rational consent to some proposed mode of treatment, nevertheless she might still be treated without kindness. For example, someone who is living in extreme poverty might receive just enough food aid not to starve, and just enough health care aid not to die from preventable causes, but also not enough aid to be well-fed, healthy, self-supporting, or able to engage in any creative, meaningful, useful, or productive activities. Then they're being oppressed (Hanna, 2018d: sections 18–20), by being condemned to a life of constant neediness and suffering.

The upshot, then, is that a human real person is being treated with sufficient respect for their human dignity if and only if (i) they're not being treated either as a mere means or as a mere thing, (ii) they can give their explicit and/or implicit rational consent to that treatment, and (iii) they're being treated with kindness. In other words, no meaningful act-intention should ever be chosen or acted

23 I'll discuss the important distinction between explicit and implicit rational consent in some detail in section 5.5 below.

upon which entails that human real persons are treated either as mere means or as mere things, without their explicit or implicit rational consent, or with cruelty. To treat a human real person without sufficient respect for their human dignity, and thus either as a mere means or as a mere thing, without their explicit and/or implicit rational consent, or with cruelty, is to harm them by violating their dignity. Therefore, it's strictly morally impermissible to harm human real persons by violating their dignity; and for the very same reasons, it's also strictly morally obligatory, to prevent or reduce dignity-violating harms to human real persons. These moral principles are also commonly known as "the negative duty not to harm" and "the positive duty to prevent harm." Equivalently, a human real person is treated with sufficient respect for their human dignity if and only if they're provided with freedom from oppression.

One direct consequence of this conception of human dignity, which overlaps significantly with early Karl Marx's political theory, as formulated, for example, in the *Economic and Philosophical Manuscripts* of 1844 (Marx, 1961; see also Fromm, 1961), is that human real persons are not commodities of any kind. Therefore, any person or social institution or system that commodifies human real persons, undermines and violates their human dignity.

Another direct consequence of this conception of human dignity, is that human real persons do not have to do anything in order to have dignity, nor can they lose their human dignity by acting badly. Human dignity is neither an achievement nor a reward for good conduct: on the contrary, it's a constitutive endowment of their human real personhood.

And another direct consequence of this conception of human dignity is that it is not a general requirement of any human real person's having dignity that they self-consciously recognize that they themselves have dignity, nor is it a general requirement of our acknowledging others as having dignity that we self-consciously recognize that they have dignity. This is for two reasons.

First, the mental act or state of recognizing oneself or another real person as having dignity is not originally or primarily an act or state of self-conscious, or reflective, report, belief, or judgment. On the contrary, it's originally and primarily an act or state of pre-reflectively conscious emotional perception, or what Michelle Maiese and I have called *affective framing* (Hanna and Maiese, 2009: section 5.3; Maiese, 2011). More precisely, on this view, emotional perception consists in an essentially embodied, conscious, feeling, desiring, passionate intentional agent's representing the world via her desire-based readiness to choose or act intentionally, and, in the midst of that readiness, being disposed to have feelings about the world, or others, or herself, in certain specific ways; and the mental content of such acts or states of emotional perception is essentially non-conceptual (Hanna, 2015a:

ch. 2). These same points are also very effectively conveyed by Wittgenstein in the *Philosophical Investigations*, without any technical terminology:

> "I believe that he is suffering."—Do I also *believe* that he isn't an automaton? It would go against the grain to use the word in both connexions. (Or is it like this: I believe that he is suffering, but am certain that he is not an automaton? Nonsense!) Suppose I say of a friend: "He isn't an automaton."—What information is conveyed by this, and to whom would it be information? To a *human being* who meets him in ordinary circumstances? What information *could* it give him? (At the very most that this man always behaves like a human being, and not occasionally like a machine.) "I believe that he is not an automaton," just like that, so far makes no sense. My attitude towards him is an attitude towards a soul. I am not of the *opinion* that he has a soul. (Wittgenstein, 1953: p. 178[e])

Second, the concept of human dignity, as I'm spelling it out in this section, is a characteristically moral-metaphysical concept that's knowable or known only by rational reflection, moral intuition, and philosophical analysis. It would be paradoxical in the extreme if, for example, someone's falling deeply in love and regarding another real person as inherently lovable required reflectively knowing the moral-metaphysical analysis of the concept of *love*, either partially or completely. On the contrary, obviously, romantic people normally affectively frame other people as inherently deeply lovable, and thereby fall deeply in love with them, without requiring any reflective or analytical grasp whatsoever of the concepts under which they themselves or the objects of their pre-reflectively conscious emotional perception fall. So too, it would be paradoxical in the extreme if, for example, someone's either being worthy of respect for their human dignity, or someone's respecting another human real person, required reflectively knowing the moral-metaphysical analysis of the concept of human dignity, either partially or completely. On the contrary, people normally affectively frame themselves and others as having dignity in a pre-reflective and non-self-consciously conscious way, and without requiring any reflective or analytical grasp whatsoever of the concepts under which they themselves or the objects of their pre-reflectively conscious emotional perception fall.

3.4 The Metaphysical Ground of Human Dignity: Human Real Personhood

The metaphysical ground of the dignity of human real persons is their real personhood. And our real personhood is an essentially embodied, unified set of innate cognitive, emotional, and practical capacities. And this is the metaphysical ground of the absolute, non-denumerably infinite, intrinsic, and objective value or worth

of all real persons, including of course all human real persons, as ends-in-themselves, precisely because this essentially embodied set of capacities is the only thing in the universe capable of freely recognizing, freely creating, freely acting according to and for the sake of, and freely sustaining, transcendental value or the highest good, which we already know to exist by the Cantorian argument sketched earlier. So human dignity is, at bottom, all about the essentially embodied complex capacity for free will and practical agency that's aimed at transcendental value or the highest good, aka free agency. But what, more precisely, is a real person?

Necessarily, every real person is also an individual animal that inherently belongs to some species or another (for example, a human real person), but the converse is not the case: not every individual animal within a species is a real person. For example, human infants born with anencephaly—without a cerebrum or a cerebellum, and lacking the top part of the skull—are really biologically human, but not human real persons. So not every individual human being is a human real person. Moreover, not every particular living organism within a species is even an individual animal within that species, much less a real person in that species. For example, normal human embryos or zygotes prior to 14 days after conception, during the period of "totipotency," are not even individual human animals, precisely because during that period they can still either split into twins or fuse with several other embryos into a *chimera* (Kuhse and Singer, 1990).

If, necessarily, all real persons are individual animals within some species or another, then obviously we can make some headway towards explicating the nature of real persons only if we are able to answer a preliminary question: "What is an animal?" *The Oxford English Dictionary* tells us that the word "animal" means "a living organism which feeds on organic matter, usually one with specialized sense organs and nervous system, and able to respond rapidly to stimuli" (Hawkins and Allen, 1991: p. 52). In the usage of contemporary biologists, the term "animal" also has a taxonomical sense, in that animals are said to constitute one of the five kingdoms of living things: Monera (bacteria), Protists, Fungi, Plants, and Animals. The class of animals that is jointly specified by these ordinary language and biological-taxonomical senses includes vertebrates and invertebrates, mammals and non-mammals—including birds, reptiles, amphibians, various kinds of fish, insects, and arachnids. My usage of the term "animal" throughout this chapter, however, is a slight precisification of the ordinary language and biological-taxonomical usages, intended also to coincide with its use in cognitive ethology, that is, the scientific study of animal minds and especially non-human animal minds in the context of macrobiology, cognitive psychology, and behavioral psychology (Griffin, 1976, 1984, 2001; Allen and Bekoff, 1997; Bearzi and Stanford, 2008). To signal this precisification, I've coined the quasi-technical term *minded animal.* Minded animals are living organisms with innate capacities for conscious-

ness, intentionality, and caring—including affective sub-capacities for desiring, emoting, feeling, and the passions.

Now minded animals are always creatures within some real species *S* or another, hence they are always *S*-type (say, human, or feline, or canine, or equine, etc.) animals. But as I noted above, not every living organism within a species *S* is an individual *S*-type animal. For example, a single human embryo or zygote (that is, the sperm-fertilized ovum) is a living organism within the human species, in the strictly phylogenetic sense of sharing our species-specific biological essence, but a single human embryo is not necessarily a human individual. This is because, as I also noted above, early human embryos up to about the 14th day of their existence are totipotent. This means, among other things, that one embryo can split and later become two distinct human individuals (twins), and also that two embryos can fuse and later become a single human individual (chimeras). What, more generally, is an individual belonging to some species S, that is, what is an individual *S*-type animal? My claim is this:

> Something *X* is an individual *S*-type (human, feline, etc.) animal if and only if *X* is a living *S*-type organism, and *X* is past the period of totipotency for that species *S*.

Within the human species—and also within a few non-human animal species—many or even most of the animals within that species can also become real persons within that species. The beginning of a real person's life for a given *S*-type animal is what I call the *neo-personhood* of that animal (Hanna, 2018c: sections 3.1–3.3). In the human species, as far as we currently know, the capacity for consciousness first manifests itself in normal fetuses between 25 and 32 weeks after conception or fertilization, hence roughly at the beginning of the third trimester (Boonin, 2003: ch. 3). My view is that this is when your very own human real personal life started—when you became a human neo-person. Prior to that, and from roughly 14 days after your parents conceived the human organism that eventually became you, there also existed a living human animal that also eventually became you—but, just like the totipotent human organism that became that human animal after 14 days, it was not yet you.

The distinction between animals within a species *S* on the one hand, and either neo-persons or actualized real persons within a species *S* on the other hand, is a deeply important difference, both metaphysically and morally. This can be looked at in two ways, with specific application to human animals. First, normal human fetuses after the period of totipotency but still before the emergence of consciousness at 25–32 weeks after conception or fertilization, are human animals *but not* human real persons, whether human neo-persons or ac-

tualized human real persons. Second, anencephalic human infants—a famous example is the real-world case of Baby Theresa (Rachels and Rachels, 2012: pp. 1–5)—are human animals, but *neither* human neo-persons *nor* actualized human real persons. Obviously these two claims, if true, will have serious implications for the morality of abortion and infanticide (Hanna, 2018c: ch. 3).

Every real person is also an *S*-type animal or living organism (but not conversely), and every individual *S*-type animal is also an *S*-type animal or living organism (but not conversely). Therefore, being an *S*-type animal or living organism (although not necessarily an individual one, in order to accommodate totipotent organisms in general and chimeras in particular) is a necessary although not a sufficient condition of real personhood. The rest of my metaphysical analysis of real personhood substantively borrows from two different sources: (i) Harry Frankfurt's hierarchical-desire theory of persons (Frankfurt, 1988b, 1988c, 1988d, 1988e), and (obviously) (ii) Kant's rationality-based theory of persons, especially as presented in *Groundwork of the Metaphysics of Morals* and the *Critique of Practical Reason.*

Frankfurt's theory of persons is based on the notion of an hierarchically-structured set of desires. The fundamental connection here is that for Frankfurt, a person is essentially identified with the constitution of her will, which in turn is a set of desires immanently structured by the capacities for rationality and free agency, and inherently governed by the norm of "decisive identification with effective first-order desires," that is, by the norm of authenticity or wholeheartedness. In a nutshell, that is my view of real persons too, although with a more explicitly and robustly Kantian twist, or rather, set of twists. But let me now explore some further specific Frankfurtian details, because they are fundamentally important for my account of real persons.

On my view a desire is a felt need for something, or a conscious going-for something. This is as opposed to an actual need for something—obviously not all felt needs are actual needs—and also as opposed to a mere pro-attitude towards something, a mere preference for something, or a mere wish for something. Frankfurt himself defines the notion of a desire somewhat more broadly, so as to include all pro-attitudes, preferences, and wishes; but in the present context, it is convenient to use my narrower and more conative notion of a desire. Desires in this sense are essentially equivalent with active, committed wants. So to desire *X* is actively and committedly to want *X*; and to desire to *X* is actively and committedly to want to *X.*

According to Frankfurt, some animals have not only what he calls *first-order desires*, which are ordinary direct desires for things, events, or real persons (for example, the infant wanting her mother), but also *effective first-order desires.* Effective first-order desires are desires that move (or will move, or would move) the

minded animal all the way to action. An effective first-order desire is the same as what Frankfurt calls a minded animal's *will* or *first-order volition.* First-order desires may or may not be accompanied by what he calls *second-order desires:* to want (not) to want *X*, or to want (not) to want to *X.* If so, then some of the second-order desires may be directed to the determination of precisely which first-order desire is to be the effective first-order desire, that is, the minded animal's will and first-order volition; and such desires are what he calls *second-order volitions.*

According to Frankfurt, whatever the order-level of desires or volitions, they can be either conscious or non-conscious. For the purposes of my discussion, however, I will concentrate exclusively on conscious desires and volitions. This is, in part, because I think that there is no such thing as a mental state, whether dispositional or occurrent, that is strictly non-conscious and not to some non-trivial degree occurrently conscious. In earlier work, Maiese and I have called this (admittedly controversial, but also, I believe, defensible) claim "The Deep Consciousness Thesis" (Hanna and Maiese, 2009: chs. 1–2; Hanna, 2011a). But in any case, and according to Frankfurt, all and only persons have second-order volitions, because all and only persons care about the precise constitution of their wills. By contrast to persons, creatures that are "wantons" have effective first-order desires, but they either lack second-order desires (hence they cannot care about the precise constitution of their wills because they lack self-conscious desires) or if they have second-order desires they nevertheless lack second-order volitions (hence even though they have self-conscious desires, they still cannot care about the precise constitution of their wills). Again, according to Frankfurt, all non-human animals, all human infants, and some human adults are wantons. Finally, for Frankfurt a person has freedom of the will if and only if she can determine, by means of a second-order volition, precisely which among her first-order desires is the effective one. This is also known as *identification* or *decisive identification* (Frankfurt, 1988b, 1988d); otherwise persons have unfreedom of the will. Wantons have neither freedom of the will nor unfreedom of the will, simply because they are not persons.

I accept much of what Frankfurt has to say about persons and their wills, and correspondingly I want to apply much of what he says to human real persons and their wills. Nevertheless, I also have a substantive disagreement with him on one mid-sized (as opposed to either major or minor) point.

My mid-sized point of substantive disagreement is that I doubt that Frankfurt's notion of personhood adequately captures the full breadth or depth of my broadly Kantian notion of human real personhood, according to which some human real persons have what I'll call *higher-level or Kantian rationality.* This, in turn, is an innate complex capacity for strict-norm-guided logical or practical

reasoning, for reflective self-consciousness, for autonomy or self-legislation, for authenticity or wholeheartedness, and for moral or non-moral responsibility. Any minded animal that also has higher-level/Kantian rationality can recognize necessary truths, judge or believe with a priori certainty, and choose or act wholeheartedly in accordance with desire-overriding non-instrumental, non-selfish, non-egoistic or non-self-interested, non-hedonistic, non-consequentialist, categorically normative reasons and duties, that is, those reasons and duties that inherently express The Categorical Imperative and the "categorical 'ought'."

By sharp contrast, what I'll call *lower-level or Humean rationality* involves only the possession of innate capacities for conscious, intentional desire-based logical or practical reasoning, for more or less momentary or occasional occurrent self-consciousness, and for self-interested, or in any case instrumental, intentional agency. Any minded animal that has lower-level rationality can recognize contingent truths, judge or believe with a posteriori certainty, and choose or act in accordance with broadly instrumental egoistic, hedonistic, or consequentialist reasons and duties, or those that express at most the "hypothetical 'ought'."

All minded animals that possess an innate capacity for higher-level/Kantian rationality also possess an innate capacity for lower-level/Humean rationality, but not the converse. For example, it's arguable that normal, healthy Great apes and perhaps also dolphins (Savage-Rumbaugh and Lewin, 1994; Griffin, 2001; Bearzi and Stanford, 2008)[24] possess an innate capacity for lower-level/Humean rationality, but not a capacity for higher-level/Kantian rationality. This is of course not to say that Great apes or dolphins are "irrational" or "non-rational" in any sense. On the contrary, it is only to say that, relative to those animals that do possess an innate capacity for higher-level/Kantian rationality, the rational capacity of Great apes and perhaps also dolphins is somewhat limited in complexity and normative power. Minded animals with a capacity for rationality in the higher-level/Kantian sense are not only constrained in their intentional agency by The Categorical Imperative or at least by some strictly universal, non-instrumental, altruistic, non-he-

24 Savage-Rumbaugh's research has been highly controversial. For an alternative view, see Tomasello and Call, 1997: esp. pp. 375–379. My own view, which I spell out and defend in (Hanna, 2018c: chs. 3–4), says that Great apes and perhaps also dolphins are non-autonomous, non-human real persons who are morally equivalent to normal human toddlers and other young children. This in turn suggests an argument strategy for those who seek to extend real-person-based legal rights to Great apes and dolphins: Since normal human toddlers and other young children clearly have real personhood and dignity, and since Great apes and (perhaps also) dolphins possess the same psychological capacities that ground real personhood and dignity, then it follows that Great apes and (perhaps also) dolphins also have real personhood and dignity, and therefore should also be accorded the same person-based legal rights. See also Siebert, 2014.

donistic, and non-consequentialist moral reasons and objective principles, they are also capable of being moved wholeheartedly by the moral emotion of respect for dignity. Or in other words, minded animals with a fully online capacity for rationality in the higher-level/Kantian sense are also capable of autonomy and what I call *principled authenticity*.

By contrast, minded animals that possess only an innate capacity for rationality in the lower-level/Humean sense are constrained in their intentional agency only by (at least some of) the axioms of rational choice theory, but not by strictly universal, non-instrumental, altruistic, non-hedonistic, and non-consequentialist moral reasons and objective principles. They are therefore not capable of autonomy or principled authenticity. Instead, they are at most capable of being moved non-authentically or non-wholeheartedly by the first-order Humean moral emotion of *sympathy* (Hume, 1978: books II and III).

Some human animals are "persons" in Frankfurt's sense, hence are human real persons in my sense, and also rational agents in the lower-level/Humean sense, and not rational agents in the higher-level/Kantian sense, but this is not because they lack the innate capacities for agency in the higher-level sense. Rather they do possess these capacities, but in the mode of real potentiality that is not yet actualized, hence it it is simply because they are not-yet rational agents in the higher-level/Kantian sense. Indeed, we—the actual and really possible readers of this book—were all of us, for a time, such creatures.

Consider, for example, normal toddlers. Normal toddlers and other young children are capable of a great many different sorts of complex affects and complex emotional states, including momentary anticipations of Kantian real personhood, but they are not (yet) higher-level rational human animals. Toddlers and other young children are therefore human real persons in the lower-level/Frankfurtian sense, but not (yet) real human persons in the higher-level/Kantian sense. Toddlers and other young children are therefore junior human real persons—as it were, tenure track assistant professors in The Realm of Ends—but not yet senior real human persons, as it were, tenured full professors in The Realm of Ends, even though they are, in terms of dignity, equal moral persons with all the other real persons. By another contrast, adolescents or semi-Kantian human real persons more generally are poised on a razor's edge somewhere between the higher and the lower levels. They are equal moral persons of the middle rank, but not card-carrying members of the classes above or below themselves—tenured associate professors in The Realm of Ends.

Thus human real personhood in the lower-level/Frankfurtian sense is a necessary and sufficient condition of human real personhood, which includes all the more or less online basic capacities of free agents, hence it entails dignity. And as we've seen, it is based on the fully online capacity for having second-order vo-

litions, which in turn contains several other distinct constituent fully online psychological capacities. Human real personhood in the higher-level/Kantian sense, i. e., moral agency, on the other hand, both includes and significantly augments human real personhood in the lower-level/Frankfurtian sense, by including the fully online capacity for principled authenticity, at least partially or to some degree. Correspondingly, human real personhood in the higher-level/Kantian sense is based on the fully online capacity for higher-level rational agency, which also contains several other distinct online psychological capacities. In order to display the internal complexity of the relationships between these capacities more fully, here's an explicit version of the two-level theory of human real personhood that I've been developing, together with my conception of human neo-personhood, in the form of a four-part metaphysical definition of human real personhood:

Part I. *X* is a *Frankfurtian human real person* ($person_f$) if and only if *X* is a human animal and *X* has fully online psychological capacities for

(i) essentially embodied consciousness or essentially embodied subjective experience,
(ii) intentionality or directedness to objects, locations, events (including actions), other minded animals, or oneself, including cognition (that is, sense perception, memory, imagination, and conceptualization), and caring (that is, desire, emotion, and feeling), especially including effective first-order desires,
(iii) lower-level/Humean rationality, that is, logical reasoning (including judgment and belief) and instrumental decision-making,
(iv) self-directed or other-directed evaluative emotions (for example, love, hate, fear, shame, guilt, pride, etc.),
(v) minimal linguistic understanding, that is, either inner or overt expression and communication in any simple or complex sign system or natural language, including ASL, etc., and
(vi) second-order volitions.

Part II. *X* is a *Kantian human real person* ($person_k$), aka *a human moral agent*, if and only if *X* is a human real $person_f$ and also has fully online psychological capacities for

(vii) higher-level/Kantian rationality, that is, categorically normative logical rationality (Hanna, 2006b: esp. chs. 6 – 7, 2006c, 2015a: ch. 5) and practical rationality, the latter of which also entails a fully online capacity for free agency, including deep freedom of the will, deep (non-)moral responsibility, autonomy (self-legislation), and wholeheartedness, hence a fully on-

line capacity for principled authenticity, at least partially or to some degree (Hanna, 2018b, 2018c).

Part III. *X* is a *human real person* if and only if *X* is either a human real person$_f$ or a human real person$_k$; and any other finite, material creature or entity *Y* is a non-person.

Part IV. If *X* is an actualized human real person, then *the neo-person* of *X* is also a human real person, where the neo-person of *X* is an individual human animal *A* that manifests the psychological capacity for consciousness and the following counterfactual is also true of *A:*

If *A* were to continue the natural course of its neurobiological and psychological development, then *A* would become *X.*

3.5 Frederick Douglass, Kant, and Human Dignity

Of course, Kant was not the only theorist of human dignity in the history of philosophy. For example, in *The Powers of Dignity: The Black Political Philosophy of Frederick Douglass*, Nick Bromell spells out the two basic ideas of Frederick Douglass's theory of human dignity:

> The first is that our exercise of our distinctive human powers not only expresses our humanity but also produces self-respect, or what Douglass now calls "natural dignity." This is why he asserts that "our natural powers are the foundation" of our rights: we have rights because we have a human worth, or dignity, that we feel deserves informal acknowledgment and formal protection. Douglass's second idea is … that our consciousness of our powers and rights depends, in some measure, on others conveying to us, through acts of respectful acknowledgment, that they, too, are conscious of and recognize our powers and rights. (As Douglass observed in an 1870 speech delivered in Washington, DC: "You know all men derive their impressions of their abilities and possibilities in some measure from the opinions of those who stand about them.") Thus, while our rights derive from our innate powers, and to that degree are inherent in our very humanity, they also rest in the hands of others, who may or may not allow us to become conscious of our powers, achieve our dignity, and exercise our rights. (Bromell, 2021: p. 43)

I completely agree with Douglass that dignity is grounded in our innate capacities, aka innate *faculties*, aka innate *powers.* That's basically a Kantian idea. Correspondingly, as we've seen, according to my broadly Kantian view, the unified set of those innate capacities, essentially embodied in a suitably complex living organism, is the human real person (Hanna, 2018b: chs. 6–7). And the human real person

has dignity, i. e., absolute, non-denumerably infinite, objective, intrinsic, irreducible, priceless value or worth, strictly by virtue of possessing those innate capacities.

Nevertheless, there are two facts about human dignity that are extremely difficult to wrap one's head around (see also section 3.3).

First, having human dignity doesn't actually require either our own self-consciousness of it, or its recognition by others, although in fact we're actually morally transformed (this is what Kant, in *Religion within the Bounds of Mere Reason*, calls the ethical "revolution of the heart" or "revolution of the will," and as Bromell points out, Douglass calls it "the sense of dignity"), when we do indeed become self-conscious of it and others do indeed recognize it.

Second, and this is perhaps what's hardest to wrap one's head around, is the fact that just as having a virtuous character and doing good things doesn't increase or establish one's human dignity, so too having a vicious character and doing bad things doesn't lessen or remove one's dignity. Indeed, even Kant nods and sometimes messes up on this one, for example, remarking in some places that people who allow themselves to be enslaved, or even voluntarily enslave themselves, have lost or forfeited their dignity. And Bromell's chapter shows that Douglass occasionally nods and messes up on this point too:

> What is that shared morality? Douglass's answer, as we shall now see, is that our moral powers, which are inherent in our very humanness, express the moral laws of the universe and establish "the laws of [our] own being." Those laws demand that we respect others and affirm their dignity if they give us signs of it. *If we disobey those laws, we forfeit our own claim to human status; we degrade ourselves.* (Bromell, 2021: p. 45, italics added)

But that's an important mistake. Kant and Douglass, alike, have temporarily confused (i) possessing human dignity and (ii) acting according to the moral principles that flow from that human dignity, thereby having sufficient respect for everyone's dignity. We're all almost always failing to choose and act in ways that sufficiently respect everyone's dignity, including our own, sometimes simply falling short of sufficient respect (a negative mode of moral badness that Augustine called the "privation of good") and sometimes outright violating people's dignity, for example, enslaving others, murder, rape, torture, and so on (positive moral badness, i. e., evil).

Relatedly, twenty or so years ago I read an article in the *New York Times* about a lawyer who specialized in defending people on death row, and she brilliantly remarked that everyone is always (non-denumerably infinitely—that's my addition)

better than the worst thing they've ever done,[25] which I interpreted as a very dignitarian insight. So, even the worst person among us, provided that their capacities haven't been irreparably damaged or destroyed by accident or disease, especially including insanity, is capable of doing good things. Indeed, sometimes very great sinners have become saints, and as Dostoevsky brilliantly pointed out through all the "sinner-saints" in his novels (Raskolnikov et al.), there's a deep psychological and moral connection between those two apparently radically different ways of being "human, all too human." This has to be right, in order to give an adequate explanation of why people are responsible for the good and bad things they do. If bad people lost their dignity, then a fortiori, their unified set of innate capacities would have gone down or have been lost, hence they couldn't act freely, hence they couldn't be responsible for the good and bad things they do, since freedom is a necessary condition of moral responsibility (Hanna, 2018c).

As Bromell describes it, Douglass's two-part conception of human dignity seems perfectly coherent and consistent to me, in relation to a broadly Kantian (even if not strictly Kant's own) conception, if we distinguish carefully between (i) the metaphysics of human dignity (which Kant also called "the metaphysics of morals"), which I've been exploring in this chapter, and (ii) the radical politics of human dignity, which includes a program for (iia) becoming authentically self-conscious of one's own human dignity and inspiring that self-consciousness in others who are currently blind to their own dignity, (iib) inspiring or provoking in people the recognition of everyone else's human dignity, especially inspiring or provoking those who are currently blind to the human dignity of people belonging to some or another oppressed group, and (iic) non-violently changing the social-institutional structure of the larger, oppressive society so that it incorporates the moral principles flowing from the metaphysical fact of human dignity. I think of Douglass and Martin Luther King Jr as two of the great geniuses of the radical politics of human dignity (see section 6.9 below). Sadly, however, Kant had little or nothing to say about the radical politics of human dignity—except for some esoteric radical texts buried in the middle of *Religion within the Bounds of Mere Reason* (see section 6.7 below)—mainly because he was seriously tempted by racist and neo-Hobbesian liberal political thinking himself, and didn't even begin to sort out those errors until the 1790s, i. e., the last decade of his life (Kleingeld, 2012: ch. 4; Hanna, 2017c, 2018e).

---

25 I did my level best to track down this article in the *New York Times* archives, without success. But I did discover that a widely-acclaimed contemporary public interest lawyer, Bryan Stephenson, also uses the same dignitarian thesis, "each of us is more than the worst thing we've ever done," as his personal motto (Equal Justice Initiative, 2024). Perhaps we read the same article.

I've often asked myself, what would a radical politics of human dignity look like in a contemporary 21st-century context? And the general answer I've come up with is this:

1. It would have to be broadly Kantian, universalist, non-sectarian, and cosmopolitan.
2. It would also have a necessary "spiritual" component of creative self-realization in the existential sense.
3. It would be radically opposed to any form of social organization based on coercion and authoritarianism.
4. It would be radically opposed to any other form of dignity-violating oppression.
5. It would also be wholeheartedly committed to designing and implementing constructive, enabling social institutions that satisfy people's true human needs.

In turn, as we'll see in chapter 6 below, I've applied this five-part general answer specifically to our engagement with and use of digital technology.

4 Human Dignity and The Highest Good: Nonideal Dignitarian Moral Theory

> A conflict of duties would be a relation between them in which one of them would cancel the other (wholly or in part).... But since duty and obligation are concepts that express the objective practical necessity of certain actions and two rules opposed to each other cannot be necessary at the same time, if it is a duty to act in accordance with one rule, to act in accordance with the opposite rule is not a duty but even contrary to duty; so a collision of duties and obligations is inconceivable. However, a subject may have, in a rule he prescribes to himself, two grounds of obligation, one or the other of which is not sufficient to put him under obligation, so that one of them is not a duty. (*MM* 6: 224)

> A foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines. (Emerson, 1957: p. 153)

> That an act qua fulfilling a promise, or qua effecting a just distribution of good, or qua returning services rendered, or qua promoting the virtue or insight of the agent, is prima facie right, is self-evident; not in the sense that it is evident from the beginning of our lives, or as soon as we attend to the proposition for the first time, but in the sense that when we have reached sufficient mental maturity and have given sufficient reflection to the proposition it is evident without any need of proof, or evidence beyond itself. It is self-evident just as a mathematical axiom, or the validity of a form of inference, is evident. The moral order expressed in these propositions is just as much a part of the fundamental nature of the universe (and, we may add, of any possible universe in which there are moral agents at all) as the spatial or numerical structure expressed in the axioms of geometry or arithmetic. In our confidence that these propositions are true there is involved the same trust in our reason that is involved in our confidence in mathematics; and we should have no justification for trusting it in the latter sphere and distrusting it in the former. (Ross, 1930: pp. 29–30)

> Hide what you have to hide
> And tell what you have to tell
> You'll see your problems multiplied
> If you continually decide
> To faithfully pursue
> The policy of truth (Depeche Mode, 1990)

4.1 Introduction

According to the broadly Kantian theory of human dignity, the essential moral implications of human dignity are captured by an hierarchically-ordered set of (either absolutely or *ceteris paribus*) universal moral principles specifying ways of always treating all human real persons with sufficient respect for their human dignity, the essence of which is the absolutely universal obligation never to treat any human real person (including oneself) as a mere means or as a mere thing,

https://doi.org/10.1515/9783111260525-007

in a thoroughly nonideal natural and social world. Moreover, human dignity and its essential moral and political implications are known by a multifaceted systematic method that includes (i) essentially reliable a priori moral intuitions of basic principles supplemented by logical rationality and reasoning, (ii) fairly reliable cognitive and practical constructive knowledge of non-basic principles under those basic principles, (iii) considered moral judgments in real-world contexts and in thought-experiments by way of applying and further specifying those basic and non-basic principles, and (iv) empathetic intersubjective moral phenomenology.

Strictly and narrowly speaking, in recent and contemporary Anglo-American professional academic philosophy, "nonideal theory" is political theorizing under the assumption that compliance to principles of justice is inherently not strict. But there is a broader and deeper sense of "nonideal theory" that is ethical theorizing under the assumption that compliance to moral principles is inherently not strict. It's this broader and deeper sense that I'm particularly interested in. More precisely, in this chapter I want to work out the basics of broadly Kantian nonideal dignitarian moral theory.

Sadly, there are nonideal worlds, and then there are nonideal worlds. How nonideal can they be? A moral contradiction, or moral dilemma, may be defined as a situation *S* in which a person *P*, given her moral principles, ought to choose or do some act *A* in *S* and also ought to choose or do some act *B* in *S*, such that *P*'s choosing or doing *A* entails her choosing or doing not-*B* in *S* and also *P*'s choosing or doing *B* entails her choosing or doing not-*A* in S. Sartre's famous case of the boy and his mother is of course one example:

> The boy was faced with the choice of leaving for England and joining the Free French Forces —that is, leaving his mother behind—or remaining with his mother and helping her carry on.... Who could help him choose? ... Who can decide a priori? Nobody. No book of ethics can tell him. The Kantian ethics says, "Never treat any person as a means but as an end." Very well, if I stay with my mother, I'll treat her as an end and not as a means; but by virtue of this very fact, I'm running the risk of treating the people around me who are fighting, as means; and conversely, if I go to join those who are fighting, I'll be treating them as an end, and, by doing that, I run the risk of treating my mother as a means. If values are vague, and if they are always too broad for the concrete and specific case that we are considering, the only thing left is to trust our instincts. That's what this young man tried to do; and when I saw him, he said, "In the end, feeling is what counts. I ought to choose whatever pushes me in one direction..." But how is the value of that feeling determined? What gives his feeling for his mother value? Precisely the fact that he remained with her. (Sartre, 2006: p. 400)

But here's an even more poignant, real-world, and, for me, close-to-home example, involving a Kantian philosopher I knew and studied philosophy with. Other things being equal, you ought to preserve rational human life; and other things being

equal, you also ought to prevent the suffering of other real persons if you can, especially the suffering of those you love most; and then your suffering sibling, parent, or life-partner lucidly asks you to assist in his or her suicide:

> The story was brief, tragic and haunting. A brilliant philosophy professor, Stephan Körner, had been found dead with his wife Edith, an NHS pioneer who had just been diagnosed as having terminal cancer. Instead of being divided by disease the couple chose to be united in death, taking a lethal overdose and breathing their last in each other's arms at their Bristol home. (Ahuja, 2000)

Moral contradictions or moral dilemmas in the sense I've just spelled out can originally derive from one or more moral principles. Sartre's case is a two-principle case, as is the Stephan and Edith Körner case. In *Sophie's Choice*, by contrast, Sophie is originally committed to a single moral principle which says that she ought to protect both of her children; yet the deeply unlucky context of action also brings it about that under her original moral principle Sophie is also committed to a subprinciple of selecting one of her children to be killed by the Nazis, since otherwise both children will be killed by them. In any case, I'll call any nonideal world in which moral contradictions or dilemmas—whether derived from one or many moral principles—do in fact occur, and also occur with alarming frequency, a thoroughly nonideal world. Our actual natural and social world, it seems clear, is a thoroughly nonideal world. Thoroughly nonideal worlds, then, are those intentional-act-worlds in which not only is it the case that compliance to moral principles is inherently not strict, but also moral dilemmas all-too-frequently happen.

Consider now a "human, all too human" real person of good will, living in our thoroughly nonideal natural and social world. When, as is almost inevitable in this world, her principles come into real conflict with one another in some morally unlucky situation, then this human real person of good will must wholeheartedly choose the lesser of several evils in that context of action, and also take complete responsibility, with no excuses, for something over which she had no control whatsoever—namely, the brute contingent fact of conflicting principles in that act-context. That brute contingent fact is also the morally tragic fact that every one of her choices in that situation will involve a violation of at least one of her principles. There is no way out. She must wholeheartedly choose, and then bravely and stoically take an awful hit.

This is what I call *The Kant-Sartre Insight* (see also Baiasu, 2011). In turn, I want to use The Kant-Sartre Insight as a rational-intuitive guide, or philosophical pole-star, to working out the basics of broadly Kantian nonideal dignitarian moral theory. In so doing, I'll develop a new interpretation and conservative extension of the highly influential and equally notorious ethical theory laid out by Kant in the *Groundwork*, the *Critique of Practical Reason*, the *Metaphysics of Morals*, and *Re-*

ligion Within the Boundaries of Mere Reason. Borrowing a famous phrase from Emerson, I call this new interpretation and conservative extension *The No-Foolish-Consistency Interpretation.* If I'm correct, The No-Foolish-Consistency Interpretation provides a unified solution to three classical problems in Kant's and Kantian ethics: (i) the problem of universalizability, or the apparent epistemic indeterminacy of tests for the generalizability and consistency of moral principles (see, e. g., O'Neill, 2006), (ii) the problem of rigorism, or the apparent over-strictness, apparent overgeneralization, and apparent overly-extended strictly universal scope, of moral principles (see, e. g., Korsgaard, 1996), and above all, (iii) the problem of moral dilemmas, or the apparent inconsistency between equally legitimate absolutely universal moral principles (see, e. g., Mason, 1996; Hill, 2002).

In solving the third of these problems, in accordance with broadly Kantian nonideal dignitarian moral theory, we'll find that Kant himself was clearly and even scandalously mistaken about the semantic structure and normative implications of a Kantian theory of moral principles, and also that W.D. Ross (Ross, 1930) was much closer to the truth about these matters, although still not quite adequate to the phenomena. Ross stood on the shoulders of Kant, and saw a little further than Kant did. My hope is that by standing, *Cirque du Soleil*-wise, on the shoulders of these two giants of Kantian ethics, and by resolutely focusing my sights on broadly Kantian nonideal dignitarian moral theory, I'll be able to see just a little further than either of them did.

4.2 The Thin Logic and the Thick Semantics of Moral Principles in Broadly Kantian Nonideal Dignitarian Moral Theory

The background conception of logic-&-semantics that I'm using in this section and throughout this book as a meta-ethical foundation, includes (i) Kant's "pure general logic" (*CPR* A50 – 57/B74 – 79; *JL* 9: 1 – 20, 91 – 150), or what we would now think of as second-order intensional monadic logic, that is, classical truth-functional logic together with a restricted predicate logic employing quantification into and also over one-place predicates only, and quantifying over individuals, i. e., first-order monadic logic (Boolos and Jeffrey, 1989: chs. 10, 22, and 25, and esp. pp. 250 – 255), but also quantifying over the Kantian concepts or finegrained intensions expressed by one-place predicates, i. e., second-order intensional monadic logic (Hanna, 2015a: ch. 5), and (ii) Kant's "transcendental logic" (*CPR* A55 – 57/ B79 – 82), or what we would now think of as a finegrained intensional possible worlds semantics of propositions together with what I call *Kantian modal dualism.*

Now according to the robust semantics of Kantian modal dualism, in turn, there are two irreducibly and essentially different kinds of necessary truth: (i) analytic necessity, which is a priori necessary truth in virtue of conceptual content, always taken together with some things in the world beyond conceptual content, although never in virtue of those worldly things, that is, the necessity that flows from concepts, and (ii) synthetic necessity, which is a priori necessary truth in virtue of things in the world beyond conceptual content, that is, truth in virtue of pure intuitional and imaginational content representing the underlying non-empirical intrinsic spatiotemporal, causal-dynamic, and mathematical immanent structures of matter in the actual world, always taken together with some conceptual content, although never in virtue of conceptual content, that is, the necessity that flows from things in the world (Hanna, 2015a: ch. 4).

So, in other words, the background conceptions of logic and semantics I'm using as a dual logico-semantic meta-ethical foundation for broadly Kantian nonideal dignitarian moral theory jointly provide, at one and the same time, for a somewhat thinner and more minimalist conception of pure logic than standard classical logic in the Frege-Russell-Carnap-Tarski-Quine tradition (aka "elementary logic") and also for a somewhat thicker and more robust conception of semantics than standard classical semantics in the mainstream Frege-Russell-Carnap-Tarski-Kripke tradition. I argued in detail and at length in *Kant the Foundations of Analytic Philosophy* and again in *Cognition, Content, and the A Priori* that although Kant is almost universally criticized for having a thinner logic and a thicker semantics than most logicians and semanticists in the mainstream Frege-Russell-Carnap-Tarski-Quine-Kripke tradition are prepared to accept, nevertheless, there are very good reasons to think that they are wrong, and Kant was right (Hanna, 2001, 2015a: esp. chs. 2, 4, and 5, 2017d). As controversial as those claims are, in order not to overburden the present book, I won't attempt to re-argue those claims here but will simply assume the soundness of my earlier arguments.

Still, I do want to emphasize right from the outset that broadly Kantian nonideal dignitarian moral theory does indeed presuppose a special non-classical logic and also a special non-classical semantics, and also that if we take this thinner logic and that thicker semantics explicitly into account, then our conception of broadly Kantian nonideal dignitarian moral theory will be significantly deepened and strengthened, as per The No-Foolish-Consistency Interpretation.

According to Kant in the *Groundwork* and also later in the *Critique of Practical Reason* (*CPrR* 5: 19–28), the *Metaphysics of Morals* (*MM* 6: 211–227), and *Religion Within the Boundaries of Mere Reason (Rel* 6: 3–5, 20–50), and also according to broadly Kantian nonideal dignitarian moral theory, morality is grounded on a set of strictly and unconditionally universal a priori normative meta-principles which are categorically binding on all rational beings, and more specifically are

categorically binding on all rational human animals, insofar as all rational human animals have (i) a "will" (*Wille*), which is an innate psychological capacity for rational desiring, or practical justification in terms of either non-instrumental reasons (namely, The Categorical Imperative in its four or five analytically equivalent formulations) or instrumental reasons, and also (ii) a "power of choice" (*Willkür*), which is an innate psychological capacity for effective desiring, or causally efficacious conscious motivation to choice and action. Otherwise put, the *Wille* is a legislative practical capacity that generates, recognizes, and is more generally reasons-sensitive to principles and imperatives, whereas the *Willkür* is an executive practical capacity that enacts and implements principles and imperatives by means of reasons-sensitive conscious conations, drives, or impulses. Together, the faculties of *Wille* and *Willkür* jointly constitute a dual faculty for rational desire-based choice. In turn, the *Wille* has (i) a higher proper part (pure practical reason, that is, a power for non-instrumental reasoning) that generates, recognizes, and reasons with categorical imperatives, and also (ii) a lower proper part (what I'll dub *impure practical reason*, that is, a power for instrumental reasoning) that generates, recognizes, and reasons with hypothetical imperatives. So, to summarize, according to my broadly Kantian approach to cognition, volition, and practical reasoning, the overall structure of the human will or faculty of desire looks like this:

Human Will or Faculty of Desire (*Begehrungsvermögen*):

higher part = faculty of practical reason or will proper (*Wille*):
    higher part = pure or non-instrumental reason
    lower part = impure or instrumental reason

lower part = power of choice (*Willkür)*

Moreover, according to the broadly Kantian nonideal dignitarian theory of moral principles that I'm developing here, we need to distinguish very sharply between (i) absolutely universal and objective moral meta-principles, which are strictly and unconditionally universal and objective a priori normative rules binding on all rational beings, including all rational human animals or real human persons, (ii) first-order substantive ceteris paribus objective moral principles (aka "fairly universal and objective moral principles"), which tell us what we ought to do, other things being equal, and are binding on all rational human animals or real human persons in any set of circumstances, provided that certain favorable background conditions obtain, and finally (iii) moral duties, which are first-order objective moral principles that are also agent-centered obligations.

Granting all that, then in given act-contexts, moral agents can find that other things really are equal. So moral duties are first-order substantive ceteris paribus objective moral principles with agent-centered application, under absolutely universal and objective moral meta-principles. A first-order substantive ceteris paribus objective moral principle is essentially the same as what Kant calls a "ground of obligation" (*MM* 6: 224). A ground of obligation is a morally sufficient reason for choosing-and-acting or for refraining, other things being equal. Similarly, to use an everyday analogy, your mother, father, or kindergarten teacher can tell you what you ought to do or not do, other things being equal, and s/he might be completely right. But because our actual natural and social world is a thoroughly nonideal world, other things really might not be equal in any given actual act-context; and, correspondingly, because the ceteris paribus condition therefore really might not be satisfied in that actual act-context, it does not automatically follow that you are obligated to do what your mother, father, or teacher rightly tells you that you ought to do—unless, in that act-context, things really are equal, and you yourself really can do it. Thus in order to be a moral duty, a first-order substantive ceteris paribus objective moral principle has to have adequate agent-centered force in an actual act-context, and this depends in part on the way the world and other people just contingently really happen to be, quite independently of the agent herself, as well as depending in part on the actual agent herself and her agential capacities in that actual act-context.

This, in turn, shows us how correctly to formulate the well-known Kantian principle that "ought implies can" (*CPR* A548/B576, A807/B835; *MM* 6:380). As Robert Stern has correctly observed, the Kantian version of "ought implies can" does *not* mean that "nothing can be right that we are incapable of achieving," but *instead* means that "we cannot be obliged to do what is right unless we are capable of acting in that way" (Stern, 2004: p. 59). As we all know, there is significant contingent variability in how our basic shared agential capacities are actually realized in different real human persons, in different act-contexts. Thus moral duties obligate us to do what some moral principles tell us we ought to do, other things being equal—that is, leaving out contingent conditions in act-contexts. But if we reintroduce contingent conditions in act-contexts, then we might be morally obligated, although it is not necessarily the case that we will be morally obligated; for we do not have a duty in each actual act-context, but rather only in some actual act-contexts. I'll come back to this point in section 4.3 below.

Necessarily, every moral duty is also a first-order substantive ceteris paribus objective moral principle, but not every first-order substantive ceteris paribus objective moral principle is also a moral duty. This is because there can be real conflicts between first-order substantive ceteris paribus objective moral principles, even in cases in which an agent has one and only one moral duty:

> A subject may have, in a rule he prescribes to himself, two grounds of obligation ..., one or the other of which is not sufficient to put him under obligation, so that one of them is not a duty. (*MM* 6: 224)

Indeed, while there can be real conflicts of first-order substantive ceteris paribus objective moral principles, there cannot be conflicts of moral duties, as a matter of analytic a priori necessity: "a collision of duties and obligations is inconceivable" (*MM* 6: 224). Thus the distinction between first-order substantive ceteris paribus objective moral principles and moral duties captures the essence of what Ross was driving at in his famous distinction in *The Right and the Good* between "prima facie duties" and "actual duties," but without the strange consequence that a given moral principle can *sort-of* be my moral duty without its also *really* being my moral duty.

In *The Right and the Good*, Ross argues that we have rational, self-evident, non-inferential, infallible a priori intuitions about an irreducibly plural class of co-basic moral principles, the *prima facie* duties, which include the seven duties of "fidelity," "reparation," "gratitude," "justice," "beneficence," "self-improvement," and "non-maleficence" (Ross, 1930: p. 21). These seven principles, purportedly, are knowable by any mature, reflective, rational human animal. *Prima facie* duties are sharply distinguished from actual duties, or duties proper, that (i) are the objectively real moral obligations binding on moral agents or persons in particular act-contexts, and (ii) are objectively determined by their being the moral principles that, in that act-context, have the greatest balance of prima facie rightness over prima facie wrongness, of all possible acts for that agent in that context, when the act is taken in "its whole nature" (Ross, 1930: pp. 20, 33, and 41). At the same time, however, according to Ross, it's not possible rationally to intuit, or authentically to know, actual duties—at best, it is possible to cognize actual duties with "right opinion," and not sufficiently justified true belief (Ross, 1930: pp. 30–31, and 145–148), that is, *essentially reliable justified true belief*, or what elsewhere I call *High-Bar justified true belief*, whereby there is an intrinsic connection between the warranting evidence for belief, as delivered by our properly functioning cognitive capacities or mechanisms, and the truth (Hanna, 2015a: section 1.2, and chs. 6–8).

In any case, Ross has three main goals in combining the classical theory of rational intuition with his theory of *prima facie* duties vs. actual duties. First, he wants to provide a secure, realistic, and a priori but also non-monistic foundation for moral theory. Second, he wants to accommodate the obvious empirical fact of conflicts of duties—moral contradictions or moral dilemmas—that seem to arise directly from the foundational fact of a plurality of basic duties together with our actual "human, all too human" existence in this thoroughly nonideal natural

and social world. And third, he wants to incorporate some measure of commonsensical or real-world fallibilism about our moral judgments in particular contexts in this world. Ross's moral intuitionism is thereby designed precisely in order to accommodate the thoroughly nonideal character of our moral lives. But his intuitionism is also philosophically notorious.

Correspondingly, here are the three classical critical objections to Ross. First, Ross's postulation of a mysterious and "queer"—in John Mackie's sense (Mackie, 1977)—faculty for intuitively knowing the prima facie duties has no independent plausibility or empirical support whatsoever. Second, Ross's infallibilism about moral intuition seems to fly in the face of the highly plausible thesis of fallibilism about a priori knowledge, as well as fallibilism about empirical knowledge. And third, the obvious empirical fact of widespread disagreement, even amongst mature, reflective rational human animals, about precisely which moral principles are true and which are false seems to undermine completely Ross's claim that even some moral principles are known intuitively with self-evidence. In addition to these three classical worries, I also have a non-classical, fourth critical objection that in certain respects is similar to John Rawls's main worry about Ross's theory, to the effect that Ross cannot ultimately avoid a theory of the lexical ordering and weighting of the supposedly equally morally binding, lexically unordered, and unweighted prima facie duties (Rawls, 1971: p. 41).[26] More precisely, my non-classical, Rawls-inspired critical objection can be posed as a dilemma:

> Ross's theory of prima facie duties, on the one hand, explicitly postulates an irreducible pluralism of basic moral principles; yet on the other hand, he implicitly presupposes a monistic deontological scale in explicating the advance from prima facie duties to actual duties—otherwise how could there be an objective determination of one moral principle's being the one which, in a given context, expresses the greatest balance of prima facie rightness over prima facie wrongness, of all possible acts for that agent in that context, when the act is taken in "its whole nature"?

In *Cognition, Content, and the A Priori*, I respond to the three classical worries about Ross's moral intuitionism, at least by implication, by developing a broadly Kantian theory of rational intuition that's an extension of a broadly Kantian theory of mathematical intuition, and also includes moral intuition as a sub-case (Hanna,

26 Rawls correctly points out that unless Ross explicitly works out a monistic objective deontological lexical ordering, or weighting scale, for his prima facie duties in relation to actual act-contexts, then his view is a dead end. My objection is simply the flip side of Rawls's worry, namely that Ross is clearly implicitly presupposing a monistic objective deontological lexical ordering or weighting scale for mapping his prima facie judgments to actual duties in actual contexts, without either admitting it or telling us precisely what it is.

2015a: chs. 6–8). In the present context, however, I want to respond directly only to the non-classical, fourth objection to Ross's moral intuitionism by extending the notion of *structuralism* from mathematics to morality. Mathematical structuralism, as an explanatory metaphysical thesis in the philosophy of mathematics (see, e.g., Benacerraf, 1965)—defended in one way, for example, by Stewart Shapiro (Shapiro, 1997, 2000: ch. 10), and in another way by Charles Parsons (Parsons, 2008: esp. chs. 3, 5–6, and 9)—says that mathematical entities (for example, numbers or sets) are not ontologically autonomous or substantially independent objects, but instead are, essentially, positions or roles in a mathematical structure, where a mathematical structure is a complete set of formal relations and operations that defines a mathematical system. What counts as an individual object of the system is thereby uniquely determined by the system as a whole. That is, any such individual object is identical to whatever possesses a specific set of intrinsic structural system-dependent properties. In a text quoted as one of the epigraphs for this section, it seems clear enough that Ross himself had a moral structuralist idea in mind:

> The moral order expressed in these propositions is just as much a part of the fundamental nature of the universe (and, we may add, of any possible universe in which there are moral agents at all) as the spatial or numerical structure expressed in the axioms of geometry or arithmetic. (Ross, 1930: pp. 29–30)

But he never systematically developed or elaborated that important thought. Nevertheless, Ross's important thought can be unpacked and effectively deployed within the framework of broadly Kantian nonideal dignitarian moral theory. So, standing on Ross's shoulders, here are the six basic ideas behind my nonideal broadly Kantian dignitarian version of moral structuralism.

First, there's a three-leveled hierarchy of moral principles, not a "flat" or nonhierarchical set of moral principles, as is usually assumed to be the case. Second, moral principles are not ontologically autonomous, substantially independent, "atomic" semantic or normative objects, but instead are, essentially, positions or roles in a moral structure, where a moral structure is a complete set of semantic relations and normative forces that defines a moral system of principles. Third, the semantic content and normative force of any individual moral principle is thereby determined by the moral system as a whole—that is, any such individual principle is identical to whatever possesses a specific set of intrinsic structural system-dependent properties. Fourth, completely convincing, intrinsically compelling, or self-evident moral intuition applies only to the top level in the hierarchy, which are procedural meta-principles, and neither to intermediate-level first-order substantive ceteris paribus moral principles, nor to bottom-level actual duties. Fifth,

the rational advance from the completely convincing, intrinsically compelling, or self-evidently intuited top-level meta-principles to the intermediate-level first-order substantive ceteris paribus principles to the bottom-level actual duties is a process of cognitive and practical construction. And finally, sixth, real conflicts of first-order substantive ceteris paribus moral principles at the intermediate level of the hierarchy are automatically resolved by a special set of level-theoretic structural constraints, taken together with one other moral meta-principle called *The Lesser Evil Principle*, which collectively fully preserve the absolutely universal objective truth and reality of the authoritatively-intuited meta-principles at the top level of the hierarchy. As the rest of this section rolls out, I'll deploy these six moral structuralist ideas against the backdrop of broadly Kantian nonideal dignitarian moral theory, in order to capture The Kant-Sartre Insight and also express the full explanatory power of the theory.

Moral principles, whether absolutely universal and objective moral meta-principles, first-order substantive ceteris paribus objective moral principles, or moral duties, should also be sharply distinguished from moral judgments, which are constructive applications of objective moral principles in particular act-contexts. Indeed, the confusion between objective moral principles and moral judgments is perhaps the most persistent fallacy in both classical and contemporary interpretations of Kant's ethics alike. It's one sort of thing to determine the logico-semantic structure and normative implications of a given objective moral principle, and another very different sort of thing to figure out in the thick of things how a given objective moral principle is to be deployed or instantiated—that is, constructively applied—in a given actual act-context.

Now, the thesis of *constructivism*, whether inside or outside moral theory, says that human minds and human agents play active, basic roles in determining and generating the content of all beliefs, truths, knowledge (especially including the knowledge of language), desires, volitions, act-intentions, and objective logical or moral principles. Correspondingly, broadly Kantian constructivism in the theory of mental content, cognition, and knowledge (aka *Erkenntnistheorie*) says that innately-specified rules essentially constrain the process by which human minds determine and generate mental representations of a manifest world that must also structurally conform to the formal constitution of their cognitive faculties (Hanna, 2001: chs. 1–2, 2006a: ch. 6, 2006b: chs. 4 and 6). And finally, broadly Kantian constructivism in moral theory says that a fundamental conception of the rational human agent essentially constrains the process by which agents determine and generate first-order substantive objective moral principles (Hill, 1989; O'Neill, 1989).

Given this backdrop, every moral judgment constructively presupposes one or more objective moral principles; but the specific character and general properties

of moral judgments cannot be automatically extended to objective moral principles, nor can the specific character and general properties of objective moral principles be automatically extended to moral judgments, since the mind-driven and agent-driven constructive process necessarily intervenes and mediates between the two. I'll come back to the important distinction between objective moral principles and moral judgments section 4.6 below. In any case, assuming for the purposes of my argument at least the intelligibility of a sharp fourfold distinction between (i) absolutely universal and objective moral meta-principles, (ii) first-order substantive ceteris paribus objective moral principles, (iii) moral duties, and (iv) moral judgments, all of which are projected into the larger theoretical frameworks of broadly Kantian constructivism and nonideal dignitarian moral theory, I want now to address the three classical problems of universalizability, rigorism, and moral dilemmas.

4.3 How to Solve the Universalizability and Rigorism Problems

In the *Groundwork*, Kant provides four (or alternatively, depending on how fine-grained one wants the theory of basic moral meta-principles to be, five) distinct formulations of The Categorical Imperative:

> *The Formula of Universal Law* (aka FUL):
>
> *Act only in accordance with that maxim through which you can at the same time will that it become a universal law.* (*GMM* 4: 421, italics in the original)
>
> [Alternative Formulation: *The Formula of the Universal Law of Nature* (aka FULN):
>
> *Act as if the maxim of your action were to become by your will a* **universal law of nature.** (*GMM* 4: 421, italics and boldfacing in the original)]

The Formula of Humanity as End-in-Itself (FHE):

> *So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means.* (*GMM* 4: 429, italics in the original)

The Formula of Autonomy (FA):

> [The] supreme condition of the will's harmony with universal practical reason, [is] the idea *of the will of every rational being as a will giving universal law.* (*GMM* 4: 431, italics in the original)

The Formula of the Realm of Ends (FRE):

> A rational being belongs as a *member* to the realm of ends when he himself gives universal laws in it but is also subject to these laws.... Do no action on any other maxim than one such that it would be consistent with it to be a universal law, and hence to act only *so that the will could regard itself as at the same time giving universal law through its maxim.* (*GMM:* 433–434, italics in the original, translation modified)

Each of the formulas of The Categorical Imperative is a procedural moral meta-principle that tells us how to select first-order moral principles. According to the broadly Kantian nonideal dignitarian theory of moral principles, there's also a lexical ordering relation between The Formula of Universal Law/Formula of the Universal Law of Nature and the other three formulas of The Categorical Imperative, considered as a single three-membered set. In other words, The Formula of Universal Law/Formula of the Universal Law of Nature is a formal presupposition of the other three procedural moral meta-principles, hence it always logically, semantically, and normatively precedes the three of them, taken as a group. More precisely, The Formula of Universal Law says that nothing will count as an objective moral principle, and in particular nothing will count as a "maxim," unless that objective moral principle or maxim consistently generalizes. Now according to Kant, a maxim is a "principle of volition" (*GMM* 4:400) or act-intention in some or another act-context. So The Formula of Universal Law, as the formal presupposition of all procedural moral meta-principles, says that nothing will as an objective moral principle, and in particular nothing will count as a morally permissible objective principle of volition or act-intention in any act-context, unless it consistently generalizes.

The Formula of the Universal Law of Nature, as I'm understanding it, is just a specification of The Formula of Universal Law, which in turn says that nothing will count as an objective moral principle, and in particular nothing will count as a morally permissible objective principle of volition or act-intention in any act-context,

unless it consistently generalizes in possible worlds that include our laws of material nature, that is, in worlds in which causality is really possible.

By contrast, the other three formulas of The Categorical Imperative are material or substantive procedural moral meta-principles. The Formula of Humanity as an End-in-Itself says that nothing will count as an objective moral principle, and in particular nothing will count as a morally permissible objective principle of volition or act-intention in any act-context, unless it essentially supports the absolute, non-denumerably infinite, intrinsic, objective value or worth of human real persons as ends-in-themselves—i. e., human dignity—by entailing that they are never treated a mere means or as mere things. The Formula of Autonomy says that nothing will count as an objective moral principle, and in particular nothing will count as a morally permissible objective principle of volition or act-intention in any act-context, unless it essentially supports the self-legislating freedom of human real persons. And finally The Formula of the Realm of Ends says that nothing will count as an objective moral principle, and in particular nothing will count as a morally permissible objective principle of volition or act-intention in any act-context, unless it essentially supports the self-legislating freedom of human real persons in a universal intersubjective community such that each human real person is considered equally or impartially in the free choices or acts of every other human real person.

Curiously, Kant says that there are "*three* ways of representing the principle of morality" (*GMM* 4: 436, italics added), namely, FUL, FHE, and FRE; but that is clearly just Homer nodding and miscounting, since he has actually provided four formulations in the immediately preceding run of text. So, charitably, what Kant really means is that there are four ways of representing "*the* principle of morality" (italics added), namely FUL, FHE, FA, and FRE. Now, granting that charitable reading, then precisely how many categorical imperatives are there? One or four? The correct answer is: both. This is because The Categorical Imperative is most correctly construed as one set of four lexically-ordered, analytically interderivable, and necessarily equivalent procedural moral meta-principles, one of which (FUL) is also the formal logical, semantic, and normative presupposition for the other three considered as a group (FHE, FA, and FRE). Each of these procedural meta-principles occupies a certain normative-semantic position in the overall structure of broadly Kantian nonideal dignitarian moral theory; each plays a certain normative-semantic role, within one and the same larger lexically-ordered, hierarchical system of broadly Kantian nonideal dignitarian moral principles; and each differs from the others only in its specific functional normative-semantic nature and in its fine-grained intensional content:

> [T]he above [four] ways of representing the [categorical imperative] are at bottom only so many formulae of the very same law, and any one of them of itself unites the other [three] in it. (*GMM* 4: 436)

And here's a directly relevant mathematical analogy. Consider the following statements, T1 to T4, four different ways of thinking about triangles. And, to make the direct relevance of the moral-mathematical analogy even more obvious, let's call the complete set of four statements *The Triangularity Imperative:*

(T1) As a geometer, you must think that triangulars are triangulars.
(T2) As a geometer, you must think that trilaterals are trilaterals.
(T3) As a geometer, you must think that triangulars are trilaterals.
(T4) As a geometer, you must think that trilaterals are triangulars.

Now statements (T1) through (T4) are all analytically interderivable and necessarily equivalent a priori truths, each of them expressing The Triangularity Imperative, but they are not synonymous. Moreover, (T1), as embedding a straight-out identity statement about triangles, is a formal presupposition of (T2) to (T4). So too, according to nonideal broadly Kantian dignitarian moral theory, The Formula of Universal Law (aka The Formula of the Universal Law of Nature), The Formula of Humanity as an End-in-Itself, The Formula of Autonomy, and The Formula of the Realm of Ends are all analytically interderivable and necessarily equivalent a priori moral principles, but they are not synonymous, and The Formula of Universal Law is a formal presupposition for the other three. Just like T1 through T4, each of the several distinct formulations of *The Categorical Imperative* is conceptually or intensionally distinct from all of the other formulations in a semantically fine-grained way. Yet at the same time they all belong to a single, multi-termed holistic conceptual network (Hanna, 2015a: section 4.6), which, in turn, is fully embedded within one and the same larger hierarchical system of principles, whether moral or mathematical. What makes this moral-mathematical analogy not merely directly relevant but also deeply relevant, is the fact that, just as (T1) through (T4), aka The Triangularity Imperative, is a single set of four analytic truths about how, as a geometer, you must think about triangles, whose subject-matter belongs to the synthetic a priori exact science of geometry, so too the four distinct formulations of The Categorical Imperative are all analytic meta-procedural principles about first-order moral principles, whose subject-matter belongs to the synthetic a priori human science (*Geisteswissenschaft*) of morality.

The broadly Kantian nonideal dignitarian theory of moral principles is not only deeply analogous to mathematics, as Ross notes: it's also deeply analogous to logic. Indeed, I've argued explicitly and at length, in *Rationality and Logic*

and *Cognition, Content, and the A Priori*, that logic and morality are essentially connected (Hanna, 2006b: ch. 7, 2015a: ch. 5). The essential connectedness of logic and morality is particularly salient when we jointly consider broadly Kantian approaches to philosophical logic and to morality alongside each other. Then it's clear and distinct that there is a significant structural analogy between (i) the logico-normative role of The Formula of Universal Law in Kant's metaphysics of morals, and (ii) the logico-normative role of The Principle of Non-Contradiction in Kant's pure general logic. The classical Principle of Non-Contradiction says that necessarily, no statement is such that both it and its negation are true. Or equivalently, in broadly Kantian terms, since Kant himself presupposes universal bivalence in pure general logic, the pure general logic version of The Principle of Non-Contradiction says that necessarily, no statement is such that it is both true and false. Hence, according to the classical and broadly Kantian pure general logic versions of The Principle of Non-Contradiction alike, there can be no "truth-value gluts" or "true contradictions." But in view of recent and contemporary work in non-classical logic, especially including dialetheic paraconsistent logic (Priest, 1987, 1998), there's good reason to reject universal bivalence. In *dialetheic* systems, truth-value gluts or true contradictions are statements that receive both classical truth-values, True and False, on some interpretations, including some theorems of logic. For example, arguably, both The Liar Sentence, which asserts its own falsity (Tarski, 1943) and also the Gödel Sentence (which asserts its own unprovability (Gödel, 1931/1967)) are true contradictions. So dialetheic systems that permit the semantic evaluation of either The Liar Sentence or the Gödel Sentence, allow for the existence of true contradictions. Dialetheic systems, in turn, are a sub-species of *paraconsistent* systems. The defining feature of a paraconsistent system is that it includes an axiom which prevents the valid derivation of every statement whatsoever from any given contradiction, a logical phenomenon which is called "Explosion." So let's call that special axiom a *No-Explosion Axiom.* By including a No-Explosion Axiom, dialetheic paraconsistent systems constrain the logical powers of contradictions in order to accommodate the possibility of true contradictions within the system, while also preventing the state of global inconsistency or complete logical anarchy or chaos, in which *every* statement is a truth-value glut.

In order to appreciate the full logico-semantic and normative force of Kant's pure general logic, from a broadly Kantian point of view, we should interpret The Principle of Non-Contradiction non-classically and in a minimal or nonideal rationally normative way, as a strictly universal, absolutely necessary, and pure a priori logical meta-principle that lays down a necessary logical constraint on what will count as a true first-order statement in any language or logical system, but also allows for the existence of true contradictions in dialetheic paraconsistent systems. My own proposal for this minimal, nonideal, strictly universal, absolutely

necessary, pure a priori, categorically normative logical principle is what I call *Minimal Non-Contradiction:*

> Accept as truths in any natural language or logical system only those statements which do not entail that it and all other statements in any or all natural languages or logical systems are both true and false.

Minimal Non-Contradiction, in turn, guarantees what I call *minimal truthful consistency.* Truthful consistency, as such, means that you must accept as truths in a natural language or logical system only those statements which do not entail that any argument in that language or system leads from true premises to false conclusions. By contrast, minimal truthful consistency means that you must accept as truths in any natural language or logical system only those statements which do not entail that *every* argument in any or all natural languages or logical systems lead from true premises to false conclusions. This latter notion of course is consistent with holding that *some* arguments in that natural language or logical system lead from true premises to false conclusions, and indeed it is also consistent with holding that some arguments in that natural language or logical system lead from the null set of premises to necessarily false conclusions. If so, then some statements in that natural language or logical system are both true and false, hence are true contradictions. So minimal truthful consistency is consistent with dialetheic paraconsistency. In other words, then, Minimal Non-Contradiction essentially secures minimal truthful consistency, and rules out Explosion. Minimal Non-Contradiction is not a strictly truth-preserving logical principle, and not even a strictly consistency-preserving logical principle—hence it's not a maximal or ideal standard of rational normativity in logic—but it nevertheless strictly rules out global inconsistency, that is, logical chaos. Logical chaos, in turn, is the ultimate result of Explosion: If every statement whatsoever in any or all natural languages or logical systems follows from a contradiction, then the negation of every statement whatsoever also follows from a contradiction, and if every statement whatsoever in any or all natural languages is both true and false, therefore every statement whatsoever in any or all natural languages or logical systems is a truth-value glut or true contradiction.

In the early 1980s, Hilary Putnam very plausibly argued that the negative version of Minimal Non-Contradiction is the one absolutely indisputable a priori truth:

> I shall consider the weakest possible version of the principle of [non-]contradiction, which I shall call the minimal principle of [non-]contradiction. This is simply the principle that *not every statement is both true and false*... [I]f, indeed, there are no circumstances in which it

> would be rational to give up our belief that *not every statement is both true and false*, then there is at least one *a priori truth.* (Putnam, 1983: pp. 100–101)

Now, Putnam and I disagree about the nature of apriority (Hanna, 2015a: ch. 7). As I see it, his view of apriority has been too heavily influenced by Quine's critique of the analytic-synthetic distinction and Quine's deflationary epistemic re-construal of the notion of the a priori. But leaving that disagreement aside, my own broadly Kantian way of making a somewhat similar point, but even more radically, is to say that Minimal Non-Contradiction *just is* The Categorical Imperative, insofar as it inherently governs all logic, cognition, science (whether formal, exact, or natural), and theorizing more generally, as rational human activities, as well as all practical and moral activities.

What, more precisely, is the connection between Minimal Non-Contradiction and The Categorical Imperative? The connection has to do with the crucial notion of cognitive and practical construction, as it's construed according to broadly Kantian constructivism. According to this construal, a specifically cognitive construction is how the human faculty of cognition (including the sub-faculties of understanding, logical reason, sensibility, and imagination) generates empirically meaningful or objectively valid judgments as outputs, given intuitions, concepts, and an actual context as inputs, under innately specified categorically normative objective principles. Thus Minimal Non-Contradiction is an innately specified, strictly universal, absolutely necessary, pure a priori, categorically normative, and immanent structural generative objective meta-principle, specifying low-bar, minimal, or nonideal rationally normative logical standards for the cognitive construction of scientific or more generally theoretical knowledge. Also according to broadly Kantian constructivism, analogously to cognitive construction, a specifically practical construction is how a person's faculty of desire produces a morally permissible causally efficacious rational choice or moral judgment—that is, a complete act-intention-implemented-in-a-context—as an output, given desires, practical reasons, and an actual act-context as inputs. Correspondingly then, according to the broadly Kantian nonideal dignitarian theory of moral principles, The Formula of Universal Law says that necessarily no moral principle, including every candidate principle of volition or act-intention, will count as a legitimate and objective moral principle unless it consistently generalizes over the entire domain of moral principles, and thereby rules out the moral equivalent of Explosion, namely moral contradictions or moral dilemmas all over the place, that is, moral chaos:

> Accept as objective moral principles, in any or all human real individual personal lives or social institutions, only those moral imperatives or ought-claims which do not entail that it and all other moral imperatives in any or all human

real personal lives or communities are both permissible (or obligatory) and impermissible.

So The Formula of Universal Law is, essentially, a minimal objective moral meta-principle of non-contradiction, that is, a higher-order or formal objective moral principle that lays down an absolutely necessary and pure a priori minimal truthful consistency and generalizability constraint on what will count as a first-order substantive ceteris paribus or material objective moral principle in the system of objective moral principles. The Formula of Universal Law tells us what the specific logical character of any first-order substantive ceteris paribus objective moral principle must be. In this way, The Formula of Universal Law is not a criterion of moral epistemology—hence it's not a super-powered first-order substantive objective moral principle. Instead, The Formula of Universal Law is a psychologically generative objective meta-principle for the practical construction of first-order substantive ceteris paribus objective moral principles and moral duties.

Therefore, The Formula of Universal Law, just like Minimal Non-Contradiction in relation to the cognitive construction of theoretical judgments, is an innately specified, strictly universal, absolutely necessary, pure a priori, categorically normative, and immanent structural generative objective meta-principle, specifying low-bar, minimal, or nonideal rationally normative moral and logico-semantic standards for the practical construction of rational choices or moral judgments. Now, all construction, as an intentional activity of rational human animals, namely human real persons, is in certain basic respects an activity of practical construction. Thus all construction, whether cognitive or otherwise, is inherently constrained and guided by The Formula of Universal Law.

But perhaps the most crucial point here is that The Formula of Universal Law says precisely nothing about how we are going to be able to apply it to particular candidates for being first-order substantive ceteris paribus material objective moral principles that are suitable for application in particular actual act-contexts. Therefore, even if it turns out to be epistemically impossible to apply The Formula of Universal Law effectively to some prospective first-order substantive ceteris paribus objective moral principles, simply because it turns out that some cases are epistemically indeterminate, this fact has no bearing on the objective validity of The Formula of Universal Law, as long as it turns out that necessarily, any actually implemented first-order substantive ceteris paribus objective moral principle has the logico-semantic property of consistent generalizability. As Kant correctly points out, the problem of epistemic indeterminacy in applying first-order substantive ceteris paribus objective moral principles belongs to moral or practical anthropology, and not to the metaphysics of morals (*GMM* 4: 388–389; *MM* 6: 217).

More specifically however, and by way of working out the rudiments of a logicizing and dignitarianly moralizing psychosemantics, just how does The Formula of Universal Law (or for that matter, The Formula of the Universal Law of Nature, The Formula of Humanity as an End-in-Itself, The Formula of Autonomy, or The Formula of the Realm of Ends) work as a psychologically generative objective meta-principle for the practical construction of first-order substantive ceteris paribus objective moral principles, and also for the construction of morally obligatory or morally permissible objective principles of volition (namely, moral duties and moral licenses)? Two points are crucial here.

First, insofar as it is applied in actual contexts, The Formula of Universal Law properly operates only on complete act-intention contents and actual or possible act-worlds. A complete act-intention content, according to broadly Kantian nonideal dignitarian moral theory, is a fully meaningful maxim. That is, a complete act-intention content is a propositional content which includes essentially non-conceptual contents—the contents of sensory intuitions, desires, and feelings, and the semantic contents of directly referential terms—and also conceptual contents in the form of an imperative indexed to an agent, plus whatever background information about that agent, other agents, or the actual world is required to make that content fully meaningful in that actual act-context. In other words, then, a complete act-intention content is the content of an-act-intention-implemented-in-an-actual-context.

Possible worlds according to my broadly Kantian view, in turn, are nothing more and nothing less than complete and mutually logically compatible sets of different conceivable ways the actual manifestly real natural and social world could have been. If you added any other concepts to one of these sets, then a contradiction would be entailed. So in other words, according to my broadly Kantian view, possible worlds are nothing more and nothing less than maximal derivability-consistent sets of concepts (Hanna, 2001: section 5.1, 2015a: sections 3.3 and 4.7). Possible act-worlds, in turn, are possible worlds in which human freedom of the will can occur, that is, possible worlds in which psychological freedom, deep freedom (which Kant calls "transcendental freedom"), and wholehearted autonomy can all occur. Since deep freedom is the naturally creative, source-incompatibilistic spontaneous power of a living, conscious, self-conscious rational human animal agent to perform basic intentional actions in the manifestly real world (Hanna, 2006a: esp. ch 8, 2009, 2018b: chs. 1–5, 2020b), then it follows that every act-world is therefore also a manifestly real natural and social world in which conscious, self-conscious intentional causation and rational human animal free agency are not only really possible, but also fully actual, natural, and social facts of life.

Second, insofar as it is applied in actual contexts, The Formula of Universal Law properly evaluates the consistent generalizability of complete act-intention

contents or fully meaningful maxims only over relevantly restricted classes of possible act-worlds—neither over all logically possible act-worlds, nor over all really or synthetically possible act-worlds, nor even over all naturally or nomologically possible act-worlds.

I'm now in a position to propose a broadly-Kantian-nonideal-dignitarian-moral-theory-driven and logico-semantically-driven solution for the classical problem of Kantian rigorism—the problem of the apparent over-strictness or apparent overly-extended strictly universal scope of Kantian moral principles. What Kant variously calls (and unfortunately misnames, due to his failing to note and heed his own crucial distinction between moral principles and moral duties) the *strict duties*, *perfect duties*, or *ethical duties* are, in fact, (i) the first-order substantive ceteris paribus objective moral principle telling us not to lie and not to make false promises (in effect, the duty to be sincere or truthful), and (ii) the first-order substantive ceteris paribus objective moral principle telling us to preserve one's own faculty for pure practical reason, not to murder other human real persons, and not to commit suicide (in effect, the negative duty not to harm human real persons). By contrast, what Kant variously calls (and again, unfortunately misnames) the *meritorious duties*, *imperfect duties*, or *duties of virtue* are, in fact, (iii) the self-regarding first-order substantive ceteris paribus objective moral principle telling us to pursue individual happiness, develop our own talents, and perfect ourselves (in effect, the duty telling us to pursue Aristotelian eudaimonia in a Kantian sense), and (iv) the other-regarding first-order substantive ceteris paribus objective moral principle telling us to promote the happiness of other human real persons, provide benefits for them, and protect them, hence treat them with kindness (in effect, the positive duty telling us to maximize public utility in a broadly Kantian sense, including providing positive goods for other real persons and also preventing harm to them, but also including an egocentrically-centered, broadly Kantian version of the Aristotelian virtue-obligation to choose and act with kindness towards others).

Granting these distinctions, then the two points I made a few paragraphs above, effectively solve the problem of Kantian rigorism by guaranteeing that no first-order substantive ceteris paribus objective moral principle—and, in particular, none of the so-called strict, perfect, or ethical duties—will ever have a strictly universal scope that is more extended than some relevantly restricted class of possible act-worlds. By the very nature of the logico-semantic evaluation of complete act-intention contents of fully meaningful maxims, then, there simply cannot be a first-order substantive ceteris paribus objective moral principle that is overstrict, overgeneralized, or overextended in its strict universality. Therefore, by the very nature of the logico-semantic evaluation of complete act-intention contents of

fully meaningful maxims, the so-called strict, perfect, or ethical duties simply cannot be applied to any cases to which they do not actually already apply.

These same points also solve the problem of what I'll call *Kantian under-rigorism*, which is that the so-called meritorious duties, imperfect duties, or duties of virtue, seem to be not strict enough, in that they do not seem to hold in all cases in which the so-called strict, perfect, or ethical duties also hold. But this problem disappears too, as soon as we realize that the so-called perfect and imperfect duties, namely, the first-order substantive ceteris paribus objective moral principles, must have exactly the same modal extension or reference, namely, they must apply to all and only the members of some relevantly restricted class of possible act-worlds. They both tell us what we ought to do in all and only such act-worlds, other things being equal. Then both the so-called perfect duties and also the so-called imperfect duties must have exactly the same modal scope, namely the total domain of possible act-worlds, under a ceteris paribus condition.

Granting that, what *is* the significant difference between the so-called perfect and imperfect duties? My proposal is that their significant difference is not at the level of modal extension or reference, since they are all first-order substantive ceteris paribus objective moral principles with the same modal scope, but instead at the level of modal intension or sense. In other words, the so-called perfect and imperfect duties are just different modes of presentation of the same class of act-worlds, that is, different egocentrically-centered and consciously-grasped aspects, or presented partitions, of the same total domain of possible act-worlds. Thus the so-called perfect duties are nothing but first-order substantive ceteris paribus objective moral principles that seem salient to the moral agent in every possible act-world, whereas the so-called imperfect duties are nothing but first-order substantive ceteris paribus objective moral principles that seem salient to the moral agent in all and only the act-worlds in which opportunities for pursuing individual happiness and perfecting oneself, or promoting the happiness of others, positively benefiting them, and preventing harm to them—hence, treating them with kindness—are also salient. But that difference in mode-of-presentation is perfectly consistent with the fact that all of the so-called perfect and imperfect duties are just first-order substantive ceteris paribus objective moral principles that apply to all and only the members of a relevantly restricted class of possible act-worlds, for any complete act-intention or fully meaningful maxim.

It's crucial to remember here, again, that the question of how anyone could ever come to know what the relevant restricted class of possible act-worlds for a given complete act-intention or fully meaningful maxim? is, is a completely separate moral-anthropological or moral-epistemological question that is simply irrelevant to the logico-semantic and normative specific character of the first-order substantive ceteris paribus objective moral principles.

4.4 How to Solve the Problem of Moral Dilemmas

At this point in my discussion, I need to make fully explicit something that I've only briefly sketched above, which is that there are three essentially different logico-semantic types of Kantian moral principles, and that these types should be carefully sorted into a three-leveled hierarchy of principles running from the highest level to the lowest level, as follows:

> LEVEL 1: Absolutely Universal and Objective Moral Meta-Principles, that is, the four analytically equivalent formulations of The Categorical Imperative, knowable by essentially reliable self-evident moral intuition.
>
> LEVEL 2: Fairly Universal First-Order Substantive Ceteris Paribus Objective Moral Principles, knowable by fairly reliable cognitive and practical construction.
>
> LEVEL 3: Moral Duties and Moral Licenses, that is, the obligatory or merely permissible complete act-intentions or fully meaningful maxims, knowable by moral judgment.

In broadly Kantian nonideal dignitarian moral theory, moral principles inherently guide the intentional acts of human real persons. So this is what I'll call *The Hierarchy Interpretation* of the broadly Kantian nonideal dignitarian theory of moral principles, not only in the sense that the principles themselves must be hierarchically sorted in this way, but also in the sense that their psychological realization in the wills of intentional agents carries the same basic hierarchical structure.

In the three-leveled hierarchy of Kantian moral principles, principles at lower levels or types presuppose all the principles at the higher levels or types. Furthermore, the principles occurring at each level are all logico-semantically and normatively equivalent with all the other principles occurring at that level.

More specifically, the various formulations of the absolutely universal and objective Categorical Imperative are all analytically equivalent with one another across logically and conceptually possible worlds at LEVEL 1. In turn, the so-called perfect and imperfect "duties," the fairly universal first-order substantive ceteris paribus objective moral principles, are all necessarily extensionally equivalent with one another across relevant, really possible act-worlds at LEVEL 2. And finally, the moral duties and moral licenses are all biconditionally equivalent with one another across the actual world at LEVEL 3.

The modal scope of each of the three levels is also distinct from the others. The absolutely universal and objective moral meta-principles at LEVEL 1, namely, the

four analytically equivalent formulations of the categorical imperatives, hold for all logically and conceptually possible worlds, or for all and only finite, embodied rational beings, that is, beings possessing an innately-specified capacity, or faculty, of pure practical reason, along with a faculty of desire and a sensibility. My own view, according to what I call The Essential Embodiment Theory of the mind-body relation and mental causation, is that necessarily, all finite, embodied rational beings are also what I call *minded animals*, and therefore living organisms (Hanna and Maiese, 2009), but this thesis isn't absolutely required for the main point I want to make here.

The main point is that the domain of finite, embodied rational beings comprehended by the moral meta-principles at LEVEL 1 is not restricted to human beings, but instead includes any finite, embodied rational beings whatsoever, whether they are human, or artificially-constructed biological systems (as are the "replicants" in Philip K. Dick's brilliant 1968 dystopian science fiction novel, *Do Androids Dream of Electric Sheep* [Dick, 1968] and in its correspondingly brilliant 1982 film version, directed by Ridley Scott, *Blade Runner*), or alien. By sharp contrast, the fairly universal first-order substantive ceteris paribus objective moral principles at LEVEL 2 hold for all and only finite, embodied rational human beings or real human persons, that is, rational beings who are also essentially embodied as living organisms, or animals, in essentially the way we are embodied, and are thereby naturally disposed to pursue and produce happiness. And finally the moral duties and moral licenses at LEVEL 3, as principles of volition or rational choice, hold for all and only the actual, context-dependent intentional acts of rational human beings or real human persons.

Figure 5 on the next page depicts a compact diagram of the structuralist system of broadly Kantian nonideal dignitarian moral principles I've just been describing.

Now according to broadly Kantian nonideal dignitarian moral theory, as a specifically nonideal dignitarian moral theory, the three-leveled hierarchy of Kantian moral principles is governed by two distinct but closely-related level-theoretic structural constraints. Here's the first structural constraint:

> **1. The No-Global-Violation Constraint.** In order for a choice or act to be permissible or obligatory in any actual act-context, there cannot be any violation of moral principles of the highest type and at the highest level in the hierarchy of principles. That is, there cannot be any violation of any of the analytically equivalent formulations of The Categorical Imperative at LEVEL l, even if there are violations of first-order substantive ceteris paribus objective moral principles at LEVEL 2.

****

The Hierarchy of Kantian Moral Principles

LEVEL 1: Absolutely Universal and Objective Moral Meta-Principles, or Analytically Equivalent Formulations of the Categorical Imperative or CI (Scope: They Apply in All Logically and Conceptually Possible Worlds, and Hold for All and Only Finite, Embodied Rational Beings)

+ The Faculty of Pure Practical Reason

which is presupposed by

The Formula of Universal Law [or The Formula of the Universal Law of Nature]

(The Formula of Universal Law corresponds to the logical principle of **Minimal Non-Contradiction**)

which is formally presupposed by

The Formulas of Humanity as an End-in-Itself, of Autonomy, and of the Realm of Ends

which are all presupposed by

LEVEL 2: Fairly Universal First-Order Substantive Ceteris Paribus Objective Moral Principles (Scope: They Apply in All Relevant, Really Possible Act Worlds, and Hold for All and Only Rational Human Beings, Other Things Being Equal)

+ Some Essential Facts about Human Nature

including

1. the need for socialization
2. the need for survival and life
3. the need for individual happiness
4. the need for pleasure and absence of pain

which are all presupposed by

the so-called "perfect duties":	the so-called "imperfect duties":
1. be sincere or truthful	3. perfect yourself
2. do not harm real persons	4. promote others' happiness, prevent harm, be kind

LEVEL 3: Moral Duties and Moral Licenses (Scope: They Apply in the Actual Act-World, i.e. This World, and Hold for All and Only Rational Human Intentional Acts)

+ All the Contingent Facts about Desires and Actual Act-Contexts

which are all presupposed by

Either

Acts with Moral Worth, Done For the Sake of the Categorical Imperative = The Good Will

Or

Acts with Moral Value but not Moral Worth, Merely Conforming to the Categorical Imperative

****

Fig. 5: The hierarchy of Kantian moral principles (Author's diagram, 2025).

The No-Global-Violation Constraint strictly forbids violations or inconsistencies of any absolute moral principles at LEVEL 1, which would be global or context-invariant violations or inconsistencies of moral principles, even though the Constraint also permits, in some act-contexts, local or context-sensitive violations or inconsistencies of first-order substantive ceteris paribus objective moral principles at LEVEL 2. And here's the second structural constraint:

> **2. The Excluded Middle Constraint.** If an agent has a moral duty in an actual act-context, then there is always one and only one moral duty for the agent in that act-context—no matter how difficult it is for the agent herself to discern it—and acting on any other conflicting first-order substantive ceteris paribus objective moral principle applying to the agent in that context, is morally impermissible in that context.

It should be obvious how The Excluded Middle Constraint relates to the sharp distinction between moral principles and moral duties. Even though several different and possibly even really conflicting first-order substantive ceteris paribus objective moral principles might apply to a given intentional agent in a given actual act-context, nevertheless she only ever has one moral duty in that context, if indeed she has any moral duty at all in that context. The Excluded Middle Constraint is a level-theoretic constraint that guarantees an empirical moral realism within the non-Platonistic, non-naturalistic (that is, not reductively or scientifically naturalistic) framework of Kantian constructivism in ethics. And it is obviously also intimately related to the fact that The Formula of Universal Law is a formally presupposed, absolutely universal and objective minimal moral meta-principle that imposes global moral consistency on the hierarchical system of moral principles in a thoroughly nonideal natural and social world.

I'm now in a position to begin to face up to the problem of moral dilemmas on behalf of the broadly Kantian nonideal dignitarian moral theory. According to the standard interpretation of the Kantian theory of moral dilemmas, the Kantian moralist accepts that there are apparent or prima facie moral dilemmas, but then denies that there are genuine or real moral dilemmas. The standard interpretation is adequately supported by the (unfortunately, very few) relevant texts in which Kant explicitly discusses moral dilemmas. But the glaring philosophical problem with this standard Kantian theory of moral dilemmas is that it just does not seem to comport with the actual facts about our moral lives on the ground in a thoroughly nonideal natural and social world. That is, everyday moral experience, moral commonsense, and above all, existential insight, all self-evidently tell us that there are real moral dilemmas, or at the very least, that moral dilemmas are really possible.

So, let's suppose instead that we accept either the actuality or the real possibility of moral dilemmas, as some existentially insightful ethicists—for example, Kierkegaard, Sartre, Bernard Williams, and Martha Nussbaum (Williams, 1973a, 1981b, 1981c; Nussbaum, 1986)—would have us do. If so, then the Kantian system of moral principles would apparently lead to contradictions or dilemmas, and "morality totters." The so-called supreme moral principle, The Categorical Imperative, would then be bogus, and pure practical reason would then be nothing

but an evolution-generated cognitive self-defense mechanism for deceiving ourselves about *the awful truth*—a morally chaotic world. Tragically, it would be even worse than Hobbes's imaginary state of nature, in which human life, although nasty and brutish, was at least mercifully short. In the world of the bogus and self-contradictory Categorical Imperative, it would be the end of the world as we morally know it, forever.

I'll call this *The Moral Doomsday Scenario.* It's clear and distinctly true that neither the actuality nor the real possibility of The Moral Doomsday Scenario can be permitted by any acceptable version of Kantian moral theory. Therefore, in order to avoid both the actuality and the real possibility of moral dilemmas, any acceptable version of Kantian morality must be able to prove that there cannot be such things as real moral dilemmas. This in turn leads to a meta-dilemma (see also Donagan, 1984, 1993; Hill, 2002): Either we accept the standard interpretation of the Kantian theory of moral dilemmas, in which case Kantian moral theory ends up not being in conformity with everyday moral experience, moral commonsense, and existential insight, by denying the actuality and real possibility of moral dilemmas; or else we accept the claims of moral experience, moral commonsense, and existential insight, in which case we leave Kantian morality open to the actuality or real possibility of The Moral Doomsday Scenario.

Of course, Kantian moral theory is not the only moral theory that has to face up to this sort of meta-dilemma (see, e.g., Sayre-McCord, 2013). More generally, for any normative ethical theory grounded on universal principles, either one rejects the existence of moral dilemmas, in which case one's meta-ethics ends up not being in conformity with everyday moral experience, moral commonsense, and existential insight, by denying the actuality and real possibility of moral dilemmas; or else one accepts the claims of moral experience, moral commonsense, and existential insight, in which case one leaves one's meta-ethics open to the actuality or real possibility of a relevant analogue of The Moral Doomsday Scenario.

Now, what is to be done? To be sure, every contemporary Kantian moral theorist has their own take on the problem of moral dilemmas (see, e.g., Korsgaard, 1996; Hill, 2002; Shiffrin, 2016: esp. ch. 1). As a defender of broadly Kantian nonideal dignitarian moral theory, however, I think that it would be a very good thing indeed if there were a fully distinct third candidate interpretation of the Kantian theory of moral dilemmas, over and above the standard interpretation and also the interpretation mandated by everyday moral experience, moral commonsense, and existential insight, that's also importantly different from those of other contemporary Kantian moral theorists. This would be a fully distinct third interpretation that is relevantly supported by the Kant texts just like the other interpretations, and independently defensible, but also resolves the meta-dilemma.

What might such a fully distinct and philosophically adequate third interpretation look like? In view of what I've already argued, one point that should immediately strike us is the obvious structural analogy between moral contradictions or dilemmas in a Kantian structuralist system of moral principles in a thoroughly nonideal natural and social world on the one hand, and alethic contradictions or paradoxes (antinomies, hypercontradictions) in classical or non-classical logical systems on the other. This in turn provides us with a philosophical working clue: What if we thought about moral dilemmas in terms of the broadly Kantian nonideal dignitarian hierarchical structuralist system of Kantian moral principles, in parallel with thinking about logical contradictions or paradoxes in non-classical logical systems? And in particular, what if we thought about a formal analogy between that hierarchical structuralist system of Kantian moral principles on the one hand, and dialetheic paraconsistent logical systems on the other?

Here's the formal analogy I have in mind. A dialetheic paraconsistent logic explicitly allows for local, logically restricted, or non-Explosive contradictions, while also explicitly ruling out global, logically unrestricted, or Explosive contradictions that lead to logical anarchy or chaos. Moreover, and now also thinking in terms of logic, semantics, and linguistic pragmatics, we can think of local, restricted, or non-Explosive contradictions as being context-sensitive and systematically variable, and of global, unrestricted, or Explosive contradictions as being context-insensitive and systematically invariant. Correspondingly then, we can hold that a properly-designed broadly Kantian nonideal dignitarian structuralist system of Kantian moral principles will explicitly allow for real local, restricted, context-sensitive and systematically variable moral dilemmas, while also explicitly ruling out global, logically unrestricted, context-insensitive and systematically invariant moral dilemmas, the moral equivalent of Explosion, which would entail moral anarchy or chaos—that is, moral contradictions or dilemmas literally all over the place: in effect, moral Doomsday. The actuality and real possibility of local moral dilemmas would conform to everyday moral experience, moral common sense, and existential insight; and ruling out global moral dilemmas would also prevent the possibility of The Moral Doomsday Scenario. So we would then have a third interpretation of the Kantian theory of moral dilemmas that prevents the meta-dilemma, and correspondingly we would also have a richer interpretation of the broadly Kantian nonideal dignitarian structuralist system of Kantian moral principles.

Now, granting me also The Hierarchy Interpretation of the broadly Kantian nonideal dignitarian structuralist system of Kantian moral principles as a working hypothesis, I can then develop it in full conjunction with the paraconsistent logic analogy, and thereby spell out—while also, as I mentioned earlier, riffing on Emerson's famous remark about consistency being the hobgoblin of small minds—what

I call *The No-Foolish-Consistency Interpretation* of our broadly Kantian nonideal dignitarian hierarchical structuralist system of Kantian moral principles, in five steps, as follows.

First, although least importantly, the relevant Kantian texts adequately support The No-Foolish-Consistency Interpretation:

> *Act only in accordance with that maxim through which you can at the same time will that it become a universal law.* (GMM 4: 421, italics in the original)

> A *conflict of duties* (*collisio officiourum s. obligationum*) would be a relation between them in which one of them would cancel the other (wholly or in part). –But since duty and obligation are concepts that express the objective practical *necessity* of certain actions and two rules opposed to each other cannot be necessary at the same time, if it is a duty to act in accordance with one rule, to act in accordance with the opposite rule is not a duty but even contrary to duty; so *a collision of duties* and obligations is inconceivable (*obligationes non colliduntur*). However, a subject may have, in a rule he prescribes to himself, two *grounds* of obligation (one or the other of which is not sufficient to put him under obligation, so that one of them is not a duty. (*MM* 6: 224, italics in the original)

On the appropriate reading of those texts, (i) the phrase "should become a universal law" means that no moral principle can be self-contradictory in a global, logically unrestricted, context-invariant way, which makes The Formula of Universal Law, when understood independently of its real-world application to particular actual act-contexts, the moral equivalent of Minimal Non-Contradiction, (ii) the phrase "a collision of duties and obligations is inconceivable" means that there cannot be global, unrestricted, context-invariant moral dilemmas, because this would entail that some moral principles are globally, unrestrictedly, and Explosively self-contradictory, (iii) the phrase "ground of obligation" means a first-order substantive ceteris paribus objective moral principle, and (iv) the phrase "a subject may have, in a rule he prescribes to himself, two grounds of obligation" means that there can still be real local, logically restricted, context-sensitive moral dilemmas between first-order substantive ceteris paribus objective moral principles.

Second, and more importantly, in order to support The No-Foolish-Consistency Interpretation in the ways I've just indicated, we need only to formulate, within the broadly Kantian nonideal dignitarian hierarchical structuralist framework of moral principles, a proper analogue of a special paraconsistent No-Explosion Axiom, that is, a moral meta-principle which explicitly allows for real local, logically restricted, context-sensitive moral dilemmas in a thoroughly nonideal natural and social world, while at the same time explicitly constraining their logico-moral powers so that logico-moral chaos cannot Explosively result from them. Correspondingly, my proposal is that we adopt the following broadly Kantian nonideal

dignitarian moral meta-principle, which, as I mentioned above, I call *The Lesser Evil Principle:*

> Given a real local moral dilemma between first-order substantive ceteris paribus objective moral principles, you ought to choose the first-order substantive ceteris paribus objective moral principle which in that actual act-context is the lesser of the several evils, in the sense that acting on it keeps rational faith with The Categorical Imperative to the greatest possible extent.[27]

By virtue of The Lesser Evil Principle, any real violations of or real inconsistencies between first-order substantive ceteris paribus objective moral principles (that is, principles at LEVEL 2 in the hierarchy), are automatically resolved by the moral agent's being required, in any given actual act-context, to choose and act on the first-order substantive ceteris paribus objective moral principle that is the lesser of several evils in that context, in the sense that this is the moral principle which in that actual context keeps rational faith with The Categorical Imperative to the greatest possible extent.

What, more precisely, do I mean by a moral principle's "keeping rational faith with The Categorical Imperative to the greatest possible extent" in a given actual act-context? One thing I mean is that a moral principle MP_1 keeps rational faith with The Categorical Imperative to the greatest possible extent in a given act-context if and only if (i) in that act-context MP_1 belongs to a holistic conceptual network which necessarily analytically includes all of the four necessarily equivalent formulations of The Categorical Imperative, as instantiated or implemented in that act-context—hence MP_1 is either analytically entailed by or analytically entails The Categorical Imperative in that act-context—and (ii) any other relevant moral principle MP_2 in that act-context is either (iia) merely analytically consistent with, or else (iib) analytically rejected by that holistic network. And another thing I mean is that a moral principle MP_1 keeps rational faith with The Categorical Imperative to the greatest possible extent in a given act-context if and only if (i) in that act-context MP_1 adequately expresses The Categorical Imperative, and (ii) any other relevant moral principle MP_2 in that act-context is either (iia) merely consistent with the adequate expression of The Categorical Imperative, or else (iib) fails adequately to express The Categorical Imperative. I'm taking these two formulations of the notion of rational-faith-keeping-to-the-greatest-possible-extent, in terms of either analytic entailment or adequate expression, to be necessarily equivalent.

---

27 For a similar idea, see Schapiro, 2006.

Third, and also importantly, we should explicitly note that insofar as The Lesser Evil Principle and The Excluded Middle Constraint together select a certain first-order substantive ceteris paribus objective moral principle to be the agent's duty in an actual act-context, all sorts of ineluctably actualist or real-world factors will contribute to determining precisely which moral principle is to be chosen, including human desires and causal act-consequences of acts, amongst which will be private utility and public utility. But it would be a serious *non sequitur* to conclude from this, that the selected principle which is my duty in that context must be an ethically egoistic or act-consequentialist principle, or that it has been selected for ethically egoistic or act-consequentialist reasons. More generally, it is crucial not to confuse the phenomenon of context-sensitivity and systematic variability in broadly Kantian nonideal dignitarian moral theory with either ethical egoism or act consequentialism. And it's equally crucial not to fall into an instrumentalist fallacy of thinking that just because every intentional act in this thoroughly nonideal natural and social world begins with human desires and causal act-consequences, therefore its guiding principle must be dependent on human desires and act-consequences. It should be obvious that there is a close parallel here with Kant's famous warning against falling into the Empiricist fallacy of thinking that just because every cognition begins with experience, therefore its content or truth must be derived from experience (*CPR* B1; Hanna, 2015a: section 4.5).

Fourth, and slightly less importantly, The Lesser Evil Principle also captures a plausible reading of this relevant Kantian text:

> When two such grounds [of obligation] conflict with each other, practical philosophy says, not that the stronger obligation takes precedence … but that the stronger *ground of obligation prevails.* (*MM* 6: 224, italics in the original)

On the appropriate reading of the text, (i) the phrase "when two such grounds [of obligation] conflict with each other" means morally unlucky real-world situations in which there is a local, logically restricted, context-sensitive moral dilemma between first-order substantive ceteris paribus objective moral principles, and (ii) the phrase "the stronger ground of obligation prevails" means that you ought to choose the first-order principle which in that context is the lesser of the several evils, in the sense that acting on it most keeps rational faith with The Categorical Imperative.

Fifth, finally, and most importantly, The Lesser Evil Principle together with The No-Global-Violation Constraint and The Excluded Middle Constraint, jointly guarantee that in any act-context in which the agent has a moral duty, there will be one and only one first-order substantive ceteris paribus objective moral principle that is her moral duty in that context. This is because the Principle to-

gether with the two Constraints collectively rule out any global, logically unrestricted, context-invariant violation or inconsistency of moral principles in the overall structuralist system of principles, even in cases of real local, logically restricted, context-sensitive violations or inconsistencies of first-order substantive ceteris paribus objective moral principles, that is, even in cases of real local moral contradictions or dilemmas.

One crucial consequence of The No-Foolish-Consistency Interpretation of the broadly Kantian nonideal dignitarian structuralist system of Kantian moral principles is that The Lesser Evil Principle reinstates, in a non-classical global version, the classical Kantian "*ought* implies *can*" principle, which would otherwise be directly threatened by the existence of real local moral dilemmas. If The Lesser Evil Principle does not hold, then if in some actual act-context I am obligated to choose or do *A* (for example, lie to a murderer because otherwise he will murder an innocent person) and also in that context I am obligated to choose or do not-*A* (that is, not lie to the murderer because it is always wrong to lie), then morally I cannot choose or do either *A* or not-*A* in that act-context—and *ought* does not imply *can.* But if The Lesser Evil Principle does hold, together with The No-Global-Violation Constraint and The Excluded Middle Constraint, then I am obligated to choose or do either *A* or not-*A* in that act-context, depending on which is contextually selected to be my duty by the Principle together with the Constraints, and then it follows that either morally I can do *A* (because in that context I am obligated to do *A*) or morally I can do not-*A* (because I am obligated to do not-*A*), but not both—and *ought* implies *can* again (see also Jaquette, 1991).

It's also crucial to notice that the process of discovering just which of the conflicting first-order substantive ceteris paribus objective moral principles is the lesser of the several evils in that actual act-context is an issue for moral judgment, but not an issue for the non-classical thin logic and the non-classical thick semantics of broadly Kantian nonideal dignitarian moral principles. The fact that it may be extremely or even almost impossibly difficult for a moral agent to discover just *which* principle is the lesser of the several evils in that actual act-context—for example, the poignant case in which you are lucidly asked by your suffering sibling, parent, or life-partner to assist in his or her suicide, as per the real-world case of Stephan and Edith Körner—is entirely irrelevant to the application of The Lesser Evil Principle to that act-context. The Lesser Evil Principle rules out the possibility of any global moral dilemma, and it thereby rules out The Moral Doomsday Scenario. But it does not in any way underestimate or undermine the force and significance either of real local moral dilemmas, or of the very real epistemic difficulties of moral judgment in these actual act-contexts, or indeed of the very real human pathos and tragedy of such situations.

As a consequence of this last point, we rightly feel deep moral pity for Sophie in *Sophie's Choice* when she must choose which of her two children will be killed by the Nazis; and we also rightly feel a deep moral terror when we imagine her awful situation. We thereby fully acknowledge the reality of local moral dilemmas, and also fully acknowledge the corresponding problem of moral judgment in local moral dilemma situations. But it would be no moral comfort whatsoever to Sophie or anyone else if The Moral Doomsday Scenario were true, if The Categorical Imperative turned out to be bogus, if moral chaos were lurking like a post-ethical Godzilla behind the thin façade of everyday etiquette and prudential conduct, and if human morality and rationality went into the abyss. There is a world of difference between introducing a "tragic sense of life" (Unamuno, 1954) into broadly Kantian nonideal dignitarian moral theory on the one hand, and being a moral terrorist like Peter Verkhovensky in Dostoevsky's *The Devils* on the other hand. Fully acknowledging the existence and implications of real local moral dilemmas is a deep and ultimately life-affirming and morality-affirming existentialist insight. But asserting or allowing for the existence and implications of real global moral dilemmas is, emphatically on the contrary, tantamount to moral nihilism, which must be eschewed and refuted by all explicit or implicit broadly Kantian nonideal dignitarians (see also Camus, 1956).

In any case, we can now see that The No-Foolish-Consistency Interpretation of the Kantian theory of moral principles is specifically designed to impose on the total structuralist hierarchy of broadly Kantian nonideal dignitarian moral principles precisely the same sort of invulnerability to global, logically unrestricted, context-invariant moral dilemmas that Tarski's meta-linguistic solution to the semantic paradoxes (and in particular The Liar Sentence) imposes on natural languages and logical systems (Tarski, 1943). Tarski's idea was to stipulate that the logical and semantic predicates, especially including the truth-predicates, for a given language, always belong to its meta-language. Then it's impossible to form a Liar sentence in any language. It's only if the hierarchy of languages and meta-languages is collapsed into a single undifferentiated "flatland" language, that paradoxes can occur. In a precisely analogous way, in the context of the broadly Kantian nonideal dignitarian structuralist system of moral principles, it's only if all moral principles were treated as if they logically, semantically, and normatively belonged to the same level, that global moral dilemmas could occur. Just to give it a convenient name, I'll call this rationally disastrous fallacy of collapsing the levels of the hierarchy of principles, whether in logic or morality, *The Flatlander Fallacy.*

In the mid-1970s, Saul Kripke noticed that, due to the irreducible presence of indexical terms in natural languages, instances of The Liar Paradox can occur contingently, and also that natural languages can consistently contain their own truth-predicates (Kripke, 1975). Five years later, Ruth Barcan Marcus also noticed that

moral dilemmas can have essentially the same logical and semantic structure as contingent occurrences of The Liar, even when sets of first-order moral principles are consistent (Marcus, 1980). Correspondingly, according to The No-Foolish-Consistency Interpretation, real context-sensitive and systematically variable conflicts of first-order substantive ceteris paribus objective principles are allowed to occur at the second level of the hierarchy of moral principles. This adequately captures the standpoint of everyday moral experience, moral commonsense, and existentialist insight. But at the same time, we are rationally enabled to live with these real local moral contradictions or dilemmas in the thoroughly nonideal natural and social world, precisely because The Lesser Evil Principle holds, moral Explosion is thereby prevented, and The Moral Doomsday Scenario is thereby ruled out.

Taken together, then, The No-Global-Violation Constraint, The Excluded Middle Constraint, and The Lesser Evil Principle collectively guarantee the intact good will of any rational "human, all too human" agent who cognitively constructs, practically constructs, and then wholeheartedly volitionally implements the broadly Kantian nonideal dignitarian hierarchical structuralist system of moral principles in our thoroughly nonideal world. Here we must always remember that, according to broadly Kantian nonideal dignitarian moral theory, as a version of Kantian constructivism, a system of moral principles is neither an abstract object in moral Platonic heaven nor a shifting bundle of merely natural facts. Instead, this Kantian constructivist system of moral principles is nothing more and nothing less than a categorically normative immanent structure in a rational human minded animal's or real human person's actual conscious and self-conscious will; and this immanent structure has both irreducible psychological reality and also robust causal efficacy.

4.5 Policy of Truth: The Murderer-at-the-Door Revisited

Now I want to apply The No-Foolish-Consistency Interpretation of the broadly Kantian nonideal dignitarian hierarchical structuralist system of moral principles to the classic example of a moral dilemma in Kantian ethics: the notorious murderer-at-the-door case, described in Kant's essay, "On a Supposed Right to Lie from Philanthropy."

As everyone who has ever taken or taught an Introductory Ethics course knows, in this classic and indeed all-too-familiar example, a murderer appears at the door of your house and demands to know whether your friend is inside (*RTL* 8: 425). The context makes it evident that he intends to kill your friend. You can either tell the truth to the murderer and let him kill your friend, or else you can prevent harm to your friend by lying to the murderer. The context

again makes it evident that you have no other relevant choices—for example, saying nothing but still somehow fooling the murderer, physically overpowering him, or calling the police in time to stop him.

What makes this example not merely notorious but indeed notoriously notorious is that, as everyone who has taken or taught Introductory Ethics also knows, in "On a Supposed Right to Lie" and elsewhere Kant himself takes the hard or purist line, and insists that you must tell the truth to the murderer and let him kill your friend. His rationale is that truth-telling or not-lying is a so-called strict, perfect, or ethical "duty":

> To be truthful (honest) in all declarations is ... a sacred command of reason prescribing unconditionally, one not to be restricted by any conveniences.... Every individual ... has not only a right but even the strictest duty to truthfulness in statements that he cannot avoid, though he may harm himself or others. (*RTL* 8: 427–428; see also *LE* 27: 59–60, 254, 257, 444–450, 604–605, 699–702)

In other words, for most contemporary philosophers, Kant comes off as a moral fanatic about not lying and truth-telling. For, quite apart from the notorious murderer-at-the-door case and relevantly similar cases, sharply contrary to Kant, almost everyone believes that (i) failing "to-tell-the-whole-truth-and-nothing-but-the-truth," and (ii) so-called "white lies," that is, according to the *Oxford English Dictionary*, "harmless or trivial lies, especially those told to avoid hurting someone's feelings," are both very often morally permissible, depending on the act-context. Even so, it remains true that, proceeding here both casuistically and with maximum interpretive charity, Kant *could* be gotten off the philosophical hook. For we *could* note that the complete title of his essay is "On a Supposed Right to Lie *from Philanthropic Motives*," which can then be taken to say that what he is rejecting is not the moral permissibility of lying as such, but only the moral permissibility of lying from philanthropic motives, i.e. from utilitarian motives. This, of course, is perfectly consistent with claiming that, provided an agent of good will had only non-selfish, non-egoistic, non-hedonistic, non-consequentialist, and thus non-philanthropic, non-utilitarian motives for lying in certain special act-contexts, then there could still be a right to lie from respect for The Categorical Imperative and for the dignity of persons—hence, a right to lie from strictly moral motives—in those special act-contexts. And in effect, that is precisely what The No-Foolish-Consistency Interpretation of the broadly Kantian nonideal dignitarian hierarchical structuralist system of moral principles will ultimately yield as a result.

Nevertheless, the historical and textual evidence for Kant's own personal moral fanaticism and purism about not lying and truth-telling really is just too overwhelming. So we must frankly admit that Kant himself is clearly and even

scandalously mistaken on this important point. He has not only forgotten or overlooked the logical, semantic, and normative implications and resources of his own theory of moral principles for solving the problem in a satisfactory way. He has also fallen headlong into the disastrous Flatlander Fallacy of forgetting or overlooking the essentially hierarchical structure of his own system of moral principles and mistakenly treating all moral principles as if they occurred at the very same level. Or in other words, on this particular point, I think that Kant badly screwed up, and also that the 1980s and 90s new wave band *Depeche Mode* were much wiser than Kant:

> Hide what you have to hide
> And tell what you have to tell
> You'll see your problems multiplied
> If you continually decide
> To faithfully pursue
> The policy of truth (Depeche Mode, 1990)

Still, as I mentioned before, I also think that by standing straddle-wise on the shoulders of Kantian ethical giants like Kant himself and Ross, we can see a little further than they did. The core of the solution provided by The No-Foolish-Consistency Interpretation, then, is that the murderer-at-the-door case, and all cases relevantly like it, are real local, logically restricted, or context-sensitive moral dilemmas, hence real conflicts between first-order substantive ceteris paribus objective moral principles. But at the same time, they do not constitute global, logically unrestricted, or context-invariant violations of the hierarchical structuralist system of moral principles—hence they are never real conflicts of duties, and thus they are never real violations of The Categorical Imperative.

So let's assume that The No-Foolish-Consistency Interpretation is correct, and also that The No-Global-Violation Constraint, The Excluded Middle Constraint, and The Lesser Evil Principle all hold. What are we to do about the murderer-at-the-door? In this act-context, each of your two options is an evil. On the one hand, if you tell the truth to the murderer and let him kill your friend, then you've allowed the murderer to harm your friend and have violated the first-order substantive ceteris paribus objective moral principle requiring us to produce positive benefits for others and prevent harm to them. But on the other hand, if you prevent harm to your friend by lying to the murderer, then you've violated the first-order substantive ceteris paribus moral principle requiring us not to lie, other things being equal. So you're between a rock and a hard place.

Nevertheless, clearly and distinctly, the lesser evil in this context is preventing harm to your friend by lying to the murderer. It is wrong to lie, other things being equal: but in this context you are morally obligated to lie. Again, it's not just mo-

rally permissible that you lie in this context: it is morally required that you lie in this context. Why? As I've just said, lying is wrong, other things being equal, and that is why this option remains an evil. But the commonsense, everyday moral phenomena of permissibly failing to-tell-the-whole-truth-and-nothing-but-the-truth and of permissible "white lies" self-evidently show that other things are very often not equal. Correspondingly, by lying in this context there's no global, logically unrestricted, or context-invariant violation of The Categorical Imperative, since you have chosen to lie solely for the sake of preventing harm to your friend, thereby stopping the murderer from treating her as a mere means or as a mere thing, like the Nazis treated people, for no good reason at all. This clearly, distinctly, and most deeply keeps rational faith with The Categorical Imperative in that act-context, by analytically entailing and adequately expressing The Formula of Humanity as an End-in-Itself. And if I'm correct that this choice clearly, distinctly, and most deeply keeps rational faith with The Categorical Imperative by analytically entailing and adequately expressing The Formula of Humanity as an End-in-Itself, the broadly Kantian nonideal dignitarian moral principle par excellence, then, since all the formulations of The Categorical Imperative are analytically equivalent, it also necessarily follows that it clearly, distinctly, and most deeply keeps rational faith with and thereby adequately expresses The Formula of Universal Law, The Formula of Autonomy, and The Formula of the Realm of Ends too.

Moreover, given The Excluded Middle Constraint, it's self-evident that in this act-context, to tell the truth to the murderer would be to accede to, or condone, the murderer's intention to harm and treat the victim as a mere means or as a mere thing, like the Nazis treated people, for no good reason at all, and thus would clearly and deeply violate The Formula of Humanity as an End-in-Itself. So in this act-context, it's self-evidently deeply morally wrong not to lie to the murderer. Telling the truth to the murderer in this act-context would be a betrayal of rational faith in The Categorical Imperative, and a failure analytically to entail and express The Categorical Imperative.

Again, it is a moral fact that you do lie in this act-context, and that lying is wrong, ceteris paribus and objectively. But lying in this context is not globally wrong: on the contrary, it's locally obligatory, and only ceteris paribus wrong. Furthermore, there are also all those commonsense, everyday permissible failures-to-tell-the-whole-truth-and-nothing-but-the-truth and so-called "white lies." All this, in turn suggests an apt and instructive, if somewhat long-winded, bumper-sticker slogan for broadly Kantian nonideal dignitarian moral guidance in a thoroughly nonideal natural and social world:

> Think Globally, Lie Locally—but Only When You Have Overriding Good Reasons to Do So, and Only For the Sake of, or at Least Consistently with, The Categorical Imperative.

And then you must bravely, stoically, and perhaps even life-affirmingly take your hit, and take complete moral responsibility, with no excuses, for your choice to lie, just like Camus's absurd existential hero Sisyphus endlessly pushing his rock up the mountain. Letting the murderer kill your friend by telling the truth, clearly and distinctly, is the massively greater evil in that act-context, and therefore you must choose the lesser evil of preventing murder by lying. Lying is chosen only as the contextually-selected means to the end of preventing murder, which in this context keeps rational faith with The Categorical Imperative to the greatest extent. But you must also take deep moral responsibility (Hanna, 2018c: esp. chs. 4–5) for your lie. That's The Kant-Sartre Insight.

This broadly Kantian nonideal dignitarian approach to the morality of lying, and The Kant-Sartre Insight, are also both confirmed by recent medical approaches to the morality of lying to dementia patients:

> When my father was declining from Alzheimer's disease, one of the things my siblings and I used to argue about was whether to correct his confusions. For example, my father, in his impaired state, expected his live-in aide to work for room and board and would lash out at her (and us) whenever he learned that she had been paid. My siblings tended to think that it was fine to lie to him about matters like this if it helped him (and us) get through one of his rancorous moods. I fought against this practice as a matter of principle. As a doctor, I had seen how even well-meaning deception, such as withholding bad news, could be damaging. To me, a healthy relationship with our father, even in his debilitated state, could be based only on truth and trust. Small lies, even if told with the best of intentions, would undermine his dignity and erode what little connection we had left with him. As it happens, the disagreement between me and my siblings mirrored a larger debate in the medical community over the past few decades about how best to treat dementia patients. Should the ethical demands of truth-telling give way to the everyday needs of dementia care? Or is telling the truth always the best way to uphold the dignity of patients and serve their mental health? Though I began the journey through my father's illness with the view that being straight with him was paramount, my experience caring for him, as well as research and reporting I did for a book, led me to change my position. *I now think that lying can be the best strategy a dementia caregiver can use, not just practically but also morally. Deception, I now believe, may uphold a different conception of the patient's dignity—respecting the integrity of his worldview, however askew it may be.... I still don't believe that we should be lying to our loved ones with dementia as a matter of course. However, I have come to learn that the relationship between ethics and treatment in dementia is a complicated one. The demands of truth-telling exist in tension with other moral imperatives, such as consolation, reassurance and empathy.* (Jauhar, 2023, italics added)

The interesting twist here is that whereas lying to the murderer-at-the-door is morally justified by sufficient respect for the human dignity of the friend who's hiding in your house, in the case of lying to dementia patients, *it's morally justified by sufficient respect for the human dignity of the person to whom you're lying.*

Here's a further point that needs to be made especially emphatically. In the act-context of lying to the murderer-at-the-door, if you could have avoided the lie and still prevented mortal harm to your friend, say, by saying nothing but still somehow fooling the murderer, or by physically overpowering him, or by calling the police in time to stop him, then you would have done so. You do not lie "from philanthropy" or for act-consequentialist reasons. On the contrary, you lie for the sake of The Categorical Imperative, and also for the sake of the human dignity of your friend, and you also tell that lie with a tragic sense of life. So your choice has moral worth, but you also take personal deep moral responsibility for the lie, with no excuses, and you bravely, stoically, and perhaps also life-affirmingly accept whatever awful hit may follow from that complete moral responsibility, whether it be in the form of agent-regret; or in the form of the moral criticism, blame, or punishment applied by to you by others; or even in the form of the sheer cruelty of others towards you. For example, in the tragically all-too-true, post-World War II, Nazi-at-the-door variant on Kant's original case, lying to the Nazi at the door means that you seriously risk going to the wall for it, that is, being cruelly murdered too, if your lie is discovered or even seriously suspected.

At the same time, it's also clearly and distinctly true that anyone else who actually criticized, blamed, or punished you for this particular lie would be a moral purist fanatic: a moral martinette, a prig, and a rule-monger.—As Kant himself personally seems to have been, in certain respects. Moreover, in the Nazi-at-the-door version, the Nazi killers who would send you to the wall for it, and cruelly murder you, would also be near-satanically evil moral monsters. Of course, the world contains all-too-many moral purist fanatics, and all-too-many near-Satanically evil moral monsters. Our world is a thoroughly nonideal natural and social world, a Wheel of Ixion, and vale of tears. It is a once-powerful and arrogant father driven to madness by tragic regret and sorrow, an eyeless man, and a cynical-wise fool, all stumbling together over a storm-blasted heath in Shakespeare's amazing *King Lear*, or in Kurosawa's equally amazing *Ran.* But that isn't your fault.

4.6 One Last Thing, by Way of Concluding This Chapter

By way of concluding this chapter, here's one last thing I've already anticipated in sections 4.3 and 4.4 above. Corresponding to The Lesser Evil Principle is a further constraint on moral judgment to the effect that a moral agent is required to

choose, in any given act-context, what seems to her to be the first-order substantive ceteris paribus objective moral principle that is the lesser of several evils in that context, in the sense that *this* is the moral principle which in that context seems to the agent to keep faith with The Categorical Imperative to the greatest possible extent. I'll call this *The Lesser Evil Principle*.*

In line with the sharp distinction I noted in section 4.2 above between objective moral principles and moral judgments, The Lesser Evil Principle* is not to be confused with The Lesser Evil Principle itself. The moral judgments of "human, all too human" moral agents, perhaps especially including philosophers, are always fallible—so, for example, obviously I could be wrong in my considered moral judgment about which principle to choose in the murderer-at-the-door case, even despite all my rational confidence in that judgment. As a matter of psychological fact, I don't seriously think that I could be wrong about that judgment—but of course I might be. Nevertheless, errors in human moral judgment are not in and of themselves errors of objective moral principle. We mustn't confuse the very real difficulties of moral judgment, and the very real human pathos and tragedy of real local moral dilemmas, with the actuality or real possibility of global moral dilemmas. That would be to commit The Flatlander Fallacy.

Moreover, errors in moral judgment also might not be something for which a moral agent can be legitimately criticized, blamed, or punished. More positively put, my broadly Kantian nonideal dignitarian moral theory-based view is that when a moral agent messes up her moral judgment, but still satisfies The Lesser Evil Principle*, then this isn't a flaw in her objective moral principles themselves, and not a lapse for which she can be legitimately criticized, blamed, or punished. She cannot be legitimately criticized, blamed, or punished for trying her hardest to select the lesser evil and keep rational faith with The Categorical Imperative. Here we must morally blame, criticize, or even punish the act that picked out the greater evil in that context, but we mustn't morally blame or criticize the character of the real person who was wholeheartedly trying to choose the lesser evil.

For example, it seems clearly and distinctly true (to me, anyhow) that the boy in Sartre's story made the morally wrong choice by staying with his mother. He should have joined the Free French Forces. The Nazis were harming and killing literally millions of innocent people and treating them as mere means or as mere things—like pieces of garbage or offal—and they had to be resisted and if possible, stopped. Here we can usefully compare and contrast the boy in Sartre's story with the story of Rick in Michael Curtiz's 1942 classic film, *Casablanca:* If Rick had stayed with Ilsa and hadn't joined the Free French Forces, then that would clearly and distinctly have been the morally wrong choice, even though it is equally obvious in the movie that Ilsa is going to be miserable without Rick, and doomed to a

loveless marriage with the noble but Karenin-esque Victor Laszlo.[28] In Sartre's case, we can morally criticize the boy's act. But I also think that we would be moral purist fanatics to criticize the boy's character for his trying wholeheartedly to choose the lesser evil.

In this way, and for the logico-semantically-driven, broadly Kantian nonideal dignitarian hierarchical structuralist moral reasons I've spelled out, it's really possible for us to live with real local moral contradictions or dilemmas in this thoroughly nonideal natural and social world. Thereby we can avoid the many small-minded hobgoblins of foolish consistency that are so rightly derided by Emerson, and also fully acknowledge Sartre's and others' deep existential insights about these morally tragic situations. Yet at the same time, we are still able to stop well short of the moral abyss of real global moral dilemmas, i. e., moral Explosion, and The Moral Doomsday Scenario. We can be defenders of full-strength broadly Kantian nonideal dignitarian moral theory, emphatically with a tragic sense of life, but also without becoming moral nihilists. That's what The No-Foolish-Consistency Interpretation is all about.

28 The force of this analogy between Sartre's case and the *Casablanca* case of course depends to some extent on how one thinks about mother-son relationships vs. romantic love-relationships vs. marriage-relationships (in short, how we think about the Anna Karenina scenario), not to mention how we fill in various important details that are not provided either by Sartre or by the many scriptwriters of *Casablanca.* But to make my point more intuitively vivid, imagine that Laszlo had been played by a relatively unattractive actor with a ponderous manner—say, Basil Rathbone or Ralph Richardson, who played Karenin in the 1935 and 1948 film versions of *Anna Karenina* respectively—and not by the more attractive and charismatic Paul Henreid.

5 Moral Principles of Dignitarian Digital Ethics

5.1 Introduction

As per chapter 1, Dignitarian Cybernetic Instrumentalism, aka DCI, as a sub-specification of Dignitarian Digital Ethics, aka DDE, says (i) that digital technology, no matter how powerful, sophisticated, or profitable, is nothing more and nothing less than a tool created by humankind for the betterment of humankind, whose use therefore can and should be strictly constrained by general and specific moral principles flowing from the concept and fact of human dignity, and (ii) that this maximally morally robust version of cybernetic instrumentalism is explicitly and fully opposed to cybernetic oppression and cybernetic totalitarianism. What I propose to do in this chapter is to formulate and defend the moral principles of DDE against the backdrop of the broadly Kantian nonideal dignitarian moral theory I worked out in chapter 4.

5.2 Four Approaches to Digital Ethics

Steve Torrance helpfully distinguishes between four approaches to ethics in general, a distinction that can in turn be extended to digital ethics in particular:

> [Here] are four approaches to ethics in general.... The four approaches have the labels *anthropocentrism*, *infocentrism*, *biocentrism*, and *ecocentrism*.... By extension these four perspectives can also be taken to offer distinctive positions on [digital ethics].... The approaches may be seen to be partially overlapping and, as a whole, far from exhaustive....
>
> Here is a sketch of each approach:
>
> – Anthropocentrism. This approach defines ethics ... as centered around human needs and interests. On this view, all other parts of the animate and inanimate world are seen as having little or no inherent value other than in relation to human goals. It goes without saying that various versions of this approach have occupied a dominant position in Western philosophy for some time. There are many variations to anthropocentrism and they are too well known to require us to go into detail here. In relation to the objectives of [digital ethics], and those of AI more broadly, the anthropocentric view sees machines, however intelligent or personlike they may be, as being nothing other instruments for human use. The project of [digital ethics] is thus interpreted, by this approach, as being essentially a kind of safety-systems engineering, or an extension of [that]. Perhaps many of those engaged in practical, narrowly focused [digital ethics] research would agree with this position.
> – Infocentrism. This approach is based on a certain view of the nature of mind or intelligence, which in turn suggests a rather more adventurous view of ethics in general and of [digital ethics] in particular. The informational view of mind, which has been extensively

https://doi.org/10.1515/9783111260525-008

discussed and elaborated within AI and cognitive science circles over many decades, affirms that key aspects of mind and intelligence can be defined and replicated as computational systems. A characteristic position within this perspective is that ethics can best be seen as having a cognitive or informational core.... Moreover, the informational/rational aspects of ethics can, it will be said, be extended to AI systems, so that we can (at least in principle) produce artificial agents that are not merely operationally autonomous, but rather have characteristics of being ethically autonomous as well.

On the infocentric view, then, machines with ethical properties would not be merely safer mechanisms, but rather ethical agents in their own right. Some proponents of the infocentric view will be sensitive to the complaint that ethical thinking and action isn't just cognitive, and that emotion, teleology, consciousness, empathy, and so forth play important roles in moral thinking and experience. They will go on, however, to claim that each such phenomenon can also be defined in informational terms and modeled in artificial cognitive systems. Even though doing so may be hard in practice, in principle it should be possible for artificial computational agents to exemplify emotion and consciousness. It would follow that such artificial agents could be (again, in principle at least) not just morally active or productive, but also morally receptive beings—that is, beings that may have their own ethical demands and rights.

Infocentrism, so defined, may be compatible with an anthropocentric approach; however, in its more adventurous forms it sees possible future AI agents as forming a new class of intelligent being whose ethical capacities and interests are to be taken as seriously as those of humans. It is also worth pointing out that infocentrism, as here understood, has a strong technocentric bias—indeed, it could also have been named "technocentrism."

- Biocentrism. On this approach, being a living, sentient creature with natural purposes and interests is at the core of what it is to have a mind, to be a rational intelligent agent. By extension, ethical status is also believed to be necessarily rooted in core biological properties. Often ethics, on this approach, is seen as having important affective and motivational elements, which, it is felt, cannot be adequately captured within an informational model; and these affective elements are seen as deriving from a deep prehuman ancestry. Adherents of this approach are thus likely to be strongly opposed to the infocentric approach. However, they may also express opposition to an anthropocentric view of ethics in that they stress the ways in which affectivity and sentience in nonhuman animals qualify such creatures to be centers of ethical concern or as holders of moral rights, even though nonhuman animals largely lack rational powers.... Biocentrism also tends toward an ethical position characterized by E.O. Wilson ... as "biophilia"—a view that claims that humans have an innate bond with other biological species and that, in view of this, natural biological processes should have an ethical privileging in our system of values, as compared with the technological outputs of modern industrial civilization and other nonliving objects in our environment. Infocentrists, by contrast, perhaps tend toward the converse ethical attitude of "technophilia."
- Ecocentrism. Just as infocentrism could be seen as one kind of natural direction in which anthropocentrism might develop, so ecocentrism could be seen as a natural progression from biocentrism. Ethics, on this approach, is seen as applying primarily to entire ecosystems (or the biosphere, or Gaia ...) and to individual organisms only as they exist within the context of such ecosystems. Ecocentrism thus opposes the anthropocentric stress on the human sphere. However, it also opposes the biocentrist's stress on the moral status of individual creatures. Particularly in its more militant or apocalyptic forms, ecocentrism

> is motivated by a sense of a growing crisis in industrial civilization coupled with a belief in the need to reject the values of technological, market-based society in favor of a set of social values that are tuned to the natural processes of the ecosphere and to the perceived threats to those processes. Thus many supporters of ecocentrism will express a deep antipathy to the technological bias that is implied by the infocentric approach.
> Despite such deep oppositions between these positions, biocentrism, ecocentrism, and infocentrism nevertheless appear to be on similar lines in at least one striking respect: Each stresses an aspiration to transcend the domain of the merely human in ethical terms. The extrahuman or posthuman emphases are rather different in the three cases: An extension of respect from the human species to a wide variety (or the totality) of other living species in the first case; an emphasis on the total natural environment (including its inanimate aspects) in the second case; and a technological, AI-based emphasis in the third case. (Torrance, 2011: pp. 119–121)

As I pointed out in chapter 1, DDE is a radically robust version of anthropocentrism, and an updated version of Kant's dignitarian ethics. As such, it's directly opposed to *infocentric* approaches to ethics. Indeed, the very idea of a Kantian ethical robot is absurd, or at the very least highly implausible (Powers, 2011).

Other varieties of anthropocentrist digital ethics include *utilitarian* approaches and *virtue ethics* approaches, where the latter is best represented, for example, by Shannon Vallor's *technomoral virtue ethics.* Vallor describes it as follows:

> [I've] identified the emerging conditions of acute technosocial opacity: rapid technological, sociopolitical, and environmental change accompanied by existential risks that make the ethical pursuit of the good life in the 21st century extraordinarily challenging and fraught with uncertainty. [I've] suggested that our current patterns of thinking about ethics and the good life may well prove ineffective, deleterious, or even catastrophic if we do not adapt them to these new technosocial realities. [I've] also claimed that new moral resources for meeting this challenge with grace and wisdom can be found in the philosophical tradition of *virtue ethics*....
>
> [I] start from the guiding concept of a virtue as a character trait that enables us to cooperatively achieve, through excellent activity, shared goods that are internal to our practices. Next, [I] recognize in the current human environment the emergence of a new kind of coordinated technosocial practice with its own distinctive moral goods. The third step is to grasp that the chances of humans flourishing in a century where conditions for life are increasingly interdependent will depend in large part on how effective we become at coordinating our decisions and activities in order to secure these moral goods. The final step is to realize that to have any hope of realizing these goods to a significant degree, we will need to begin to cultivate in ourselves that set of moral virtues most likely to foster their realization.
>
> *Our* present moral context, the one to which our virtues must be more effeectively adapted, is one of increasingly rapid, transformative, global, unpredictable, and interdependent technosocial change. Of course the virtues needed to flourish in these conditions cannot simply replace local and settled conceptions of human excellence, for it is unreasonable to think that we can live well globally without also flourishing in our own families, states, and cultures....
> The twelve technomoral virtues that shall be [my] focus are:

1. Honesty
2. Self-Control
3. Humility
4. Justice
5. Courage
6. Empathy
7. Care
8. Civility
9. Flexibility
10. Perspective
11. Magnanimity
12. Technomoral Wisdom (Vallor, 2016: pp. 23, 50–51, 119–120)

To be sure, I don't disagree with Vallor's identification of the fundamental problem of digital ethics, namely, what she calls *acute technosocial opacity:*

> rapid technological, sociopolitical, and environmental change accompanied by existential risks that make the ethical pursuit of the good life in the 21st century extraordinarily challenging and fraught with uncertainty. (Vallor, 2016: p. 23)

Instead, my worries flow principally from the fact that Vallor's technomoral virtue ethics is a sub-species of *contemporary virtue ethics.*

According to contemporary virtue ethics, a virtue is a trait of character, manifested in habitual action, that it's good for a person to have. This is, on the whole, the same as Aristotle's virtue ethics. But there are two non-trivial differences between Aristotle's virtue ethics and contemporary virtue ethics that are worth pointing up.

First, Aristotle's virtue ethics is explicitly a version of *eudaimonism*, i. e., the morality of happiness; but contemporary virtue ethics is not necessarily a morality of happiness. Contemporary virtue ethics focuses primarily on morally good character or morally good personality, not on happiness. Happiness may quite commonly or even naturally flow from a morally good character or personality, but it need not necessarily do so.

Second, Aristotle's virtue ethics is somewhat unclear about the role of rational principles in choice, whereas contemporary virtue ethics is explicitly committed to what is sometimes called *moral particularism*, which is the rejection of both the need for and also the intelligibility of general moral principles, a view that is sometimes called *moral generalism* (see also Shafer-Landau, 2015: ch. 15). More specifically, according to moral particularism, there are only good or bad moral judgments in particular historical contexts and particular social communities, not good or bad general moral principles abstracted from contexts or communities. So if contemporary virtue ethics is true, then moral generalism is false. It's important

to note in this connection that contemporary virtue ethics, via its moral particularism, heavily relies on the idea that moral communities, and their commonly-held or standing moral beliefs, are the ultimate sources of the justification of the good and bad moral judgments we make in particular historical contexts. So contemporary virtue ethics is also a version of moral communitarianism. It's also important to note that any version of moral generalism that also asserts the existence of absolutely universal objective moral principles—so that it's, as it were, moral generalism with an attitude—is an absolutist moral generalism. Broadly Kantian dignitarian ethics is committed to absolutist moral generalism. So if contemporary virtue ethics, of which Vallor's technomoral virtue ethics is a sub-species, is true, then broadly Kantian dignitarian ethics is false, and since DDE is a sub-species of broadly Kantian dignitarian ethics, then DDE is also false; conversely, if broadly Kantian dignitarian ethics and DDE are true, then contemporary virtue ethics and Vallor's technomoral virtue ethics are false. In other words, DDE and Vallor's technomoral virtue ethics are logical contraries.

Contemporary virtue ethics provides a moral framework and moral theory based on three basic ideas: (i) that the highest human good is a morally good character or personality, (ii) that therefore we ought to be morally virtuous, and (iii) that moral virtue is expressed or shown by a person's making good moral judgments against the necessary backdrop of a particular social community in a particular historical context. But there are at least four basic critical worries about this moral framework and theory.

First, there's what James Rachels aptly calls "the problem of incompleteness" (Rachels and Rachels, 2012: pp. 170 – 171). This problem emerges when we ask: *why* are virtuous choices and actions morally good and right? If you answer, "because that's just what virtuous people do," then you've begged the question. So there must be a theory of good/bad and right/wrong that's rationally justified independently of the fact of moral virtue itself. Indeed, it's hard to imagine any non-question-begging answer to the question of why virtuous choices and actions are morally good and right, that's not ultimately an appeal to either (i) (e. g., Utilitarian) consequentialism and its benefits-based reasoning or (ii) (e. g., broadly Kantian dignitarian) non-consequentialism and its intrinsic-value based reasoning. Therefore, moral virtue is at best a derivative and secondary moral fact, not a primitive and primary moral fact.

Second, there's what I'll call *the too few virtues problem.* This problem emerges when we notice that there are many more particular historical contexts in which the morality of choices or intentions and acts is a burning issue, than there are specific moral virtues to deal with them. Aristotle and some contemporary virtue ethicists finesse this problem by appealing to practical wisdom, which can project beyond moral virtues somewhat. But if practical wisdom is doing all the basic

moral work, then virtue ethics is primarily a rationality-based moral theory, not primarily a virtue-based theory.

Third, there's what I'll call *the unity of the virtues problem*. This problem emerges when we ask: "is it possible to have one moral virtue without having all of them?" Virtue ethics is committed to a holistic thesis of the unity of the virtues: to have one moral virtue is to have all of them, that is, it is to have the entire package of moral virtues. But on the contrary, it seems really possible for someone to be morally virtuous in some respects, but not in others, and thus not be all-around morally virtuous. So, for example, it seems really possible that a generous or tolerant person might also be a coward. If this is so, then even having lots of moral virtues doesn't entail that you're an all-around morally virtuous person, and something other than merely possessing some virtues—for example, practical wisdom—must account for the unity of the virtues. But again, if practical wisdom is doing all the basic moral work, then again virtue ethics is primarily a rationality-based moral theory, not primarily a virtue-based theory.

Fourth, finally, and most problematically of all, there's what I'll call *the problem of bad and immoral virtues*. This problem emerges when we notice that Aristotelian virtues are supposed to be shared across all rational human animals. But according to contemporary virtue ethics, every moral virtue must include an essential reference to a particular social community in a particular historical context. But because this is true, it seems really possible that a particular social community can be generally wicked—for example, a slave-owning culture, like Plato's and Aristotle's Athens, or 18th and 19th-century America—and yet individual people in that particular social community in that particular historical context would still count as morally virtuous internally and relatively to that community and that context. Consider, for example, Plato and Aristotle; or Washington and Jefferson. More specifically, is it really possible to be a morally virtuous slave owner, internally and relatively to that person's particular social community in that particular historical context? Or even more forcefully put, is there any rational doubt whatsoever that Washington and Jefferson were indeed both morally virtuous slave-owners in 18th and 19th-century America? If so, then moral virtue doesn't entail moral goodness or rightness, and therefore the fact of moral virtue cannot be the foundation of morality. Since Vallor's technomoral virtue ethics is a sub-species of contemporary virtue ethics, then it automatically inherits all four of these problems.

By sharp contrast to Vallor's technomoral virtue ethics, which is indeed a serious competitor to DDE, I think that utilitarian digital ethics (see, e. g., Grau, 2011) is an obvious non-starter, in view of a host of classical worries about utilitarianism in general, and about act-utilitarianism and/or rule-utilitarianism more specifically (see, e. g., Williams, 1973b; Rachels and Rachels, 2012: chs. 7–8; Shafer-Landau, 2015: chs. 9–10; Hanna, 2018b: ch. 1).

In any case, and most importantly, massively most approaches to digital ethics are infocentric. Leading infocentrists include the following: Agar, 2020; Allen, Wallach, and Smit, 2011; Anderson and Anderson, 2011; Bostrom, 2014; Bostrom and Yudkowsky, 2014; Bryson, 2010, 2019; Calvarly, 2011; Clarke, 2011; Coeckelbergh, 2020; Dietrich, 2011; Floridi, 2011, 2013, 2016; Gips, 2011; Gordon and Nyholm, 2021; Grau, 2011; Guarini, 2011; Gunkel, 2012; Hall, 2011; McDermott, 2011; Loh, 2019; Mackworth, 2011; Marcus and Davis, 2020; McLaren, 2011; Misselhorn, 2018; Moor, 2011; Müller, 2023; Bringsjord et al., 2011; Seville and Field, 2011; Shadbolt and Hampson, 2024; Sullins, 2011; Torrance, 2011; Turilli, 2011; Whitby, 2011. My critical reply to all of them, *en masse*, can be found in detail in chapter 2, but can also be reprised here in a nutshell: if, in view of the nine arguments I provided against the strong AI thesis, I'm correct that AI really is a myth, then infocentrist digital ethics is false from the get-go.

As to biocentrism and ecocentrism, to the extent that they're both transhumanist, then they fall under my critique of transhumanism in section 2.2 above. But to the extent that they're both organicist, then they're coherent and compatible with DDE.

5.3 Our Moral Obligations with Respect to Humanity

In any case, having stated DDE's radically robust anthropocentric commitments clearly and explicitly, and having briefly compared and contrasted it with infocentrism, biocentrism, and ecocentrism, I'll now describe and defend the basic details of DDE.

All human real persons have a fundamental need for happiness, aka flourishing, and happiness or flourishing consists in the satisfaction of what I call *true human needs*. True human needs are opposed to merely self-perceived and false human needs. It might be that someone perceives within themselves an intense need to own a certain luxury automobile, even though they already own a car that is perfectly adequate to their true human needs. Therefore, it is not unkind of us not to cater to this self-perceived and false human need.

Some true human needs are such that their active satisfaction is a necessary condition of all human dignity. I'll call those *the lower-level true human needs*. For example, among the lower-level true human needs are everyone's needs for (i) adequate nourishment, adequate clothing, and adequate accommodation (provision), (ii) adequate physical and mental health, as sustained by adequate healthcare, (iii) adequate access to a healthy natural environment, both local and global, (iv) adequate scope for human movement and travel across the earth, (v) adequate protection from coercion by others (safety), (vi) adequate access to human communica-

tion and human interaction, and (vii) adequate primary and secondary education for the development and exercise of the innate capacities that collectively constitute personhood. By "adequate" in each case, I mean sufficient, in view of all relevant empirically well-supported tests that also fully conform to basic moral principles of human dignity. Since satisfying these true human needs is a necessary condition for human dignity, then sufficient respect for human dignity demands that everyone, everywhere should always have enough of whatever it takes to satisfy their lower-level true human needs.

Over and above the lower-level true human needs, all other true human needs are those whose satisfaction most fully conform to the absolute, non-denumerably infinite, intrinsic, objective value of human dignity. Indeed, they are humanity-realizing needs. More precisely, the satisfaction of such needs allows people to activate and to exercise their various capacities and realize their potentiality for being autonomous, for individually flourishing, and for collectively flourishing, in ways that also are fully compatible with and fully supportive of the agential autonomy, relational autonomy, individual flourishing (i. e., personal happiness), and collective flourishing (i. e., interpersonal or social happiness) of everyone else. Let's call these *the higher-level true human needs*, since they presuppose the satisfaction of lower-level basic human needs, their satisfaction most fully realizes the innate capacities of our human real personhood, and their satisfaction brings about individual and collective human flourishing. For example, among the higher-level basic needs are everyone's needs for (i) aesthetic enjoyment of all kinds, (ii) personal relationships of all kinds, for example, families, life-partners, lovers, close friends, a wider circle of friends, comrades, etc., (iii) social and political solidarity of all kinds, (iv) free thought and free speech of all kinds, (v) creative self-expression of all kinds, (vi) meaningful work of all kinds, (vii) higher education of all kinds, and (vii) spirituality of all kinds. Since it's arguable that the ultimate goal, purpose, or meaning of human life is no more and no less than to pursue the satisfaction of higher-level true human needs, then sufficient respect for human dignity also demands that everyone, everywhere, should always have enough of whatever it takes for them to be able to pursue their higher-level true human needs (Maiese and Hanna, 2019: ch. 3).

According to broadly Kantian nonideal dignitarian moral theory, it's morally obligatory for everyone to treat all human real persons, including ourselves, with sufficient respect for their/our dignity, which in turn entails (i) never treating human real persons as mere means to our own goals, and especially never treating them as mere things (especially via coercion), (ii) always treating human real persons consistently and impartially (except insofar as an expression of partiality—say, towards one's own family members—also sufficiently respects their dignity as well as the dignity of others), (iii) always treating human real persons in a

way that makes it really possible for them to be happy, i. e., to flourish, and therefore to satisfy their lower-level human needs and their higher-level true human needs, alike, (iv) never treating human real persons in such a way as to undermine their capacity for free agency, and (v) always treating human real persons as belonging to a universal, worldwide community of human real persons who mutually sufficiently respect one another's human dignity.

Any violation of any of the five moral obligations just formulated, whether it's aimed by a given human real person at another individual or group of individuals specifically, or whether it is built into the very structure of a social institution, is what I'll call *oppression*; and more specifically, I'll call the former kind of oppression *individual oppression*, and the latter kind of oppression *systemic oppression.*

5.4 Moral Obligations with Respect to the Design, Creation, and Use of Digital Technology

It's morally impermissible to design, create, and/or use digital technology in such a way as to fall short of, or more generally violate, sufficient respect for the human dignity of human real persons. This covers an extremely wide variety of impermissible designs, creations, and/or uses, including: (i) digital manipulation of human real persons via "nudging" strategies specifically designed to undermine the satisfaction of true human needs, (ii) digital demagoguery, (iii) digital coercion of human real persons, (iv) digital threats of violence or other significant nonviolent harm directed at human real persons more generally, (v) digital slander of individual human real persons, (vi) active digital discrimination of any kind whatsoever against human real persons of any kind whatsoever, (vii) algorithms that prevent people from gaining access to digitally-stored data or information that's necessary for being treated with sufficient respect for their human dignity, especially including making it really possible for them to satisfy of lower-level and higher-level true human needs (i. e., their individual and collective flourishing or happiness), (viii) algorithms that prevent people from gaining access to digitally-stored data or information that's necessary for their self-protection against any sort of individual oppression or systemic oppression, (ix) algorithms that prevent people from gaining access to digitally-stored data or information that's necessary for their ability to protect others against any sort of individual oppression or systemic oppression, (x) algorithms that prevent people from gaining access to digitally-stored data or information that's necessary for their ability to push back against any sort of individual oppression or systemic oppression, (xi) legally-permitted ubiquitous digital data or information collection without rational consent (see also section 5.5),

and (xii) any kind of digital technology that systematically shapes our lives in bad, false, and wrong ways (see chapter 6 below).

In this connection, it's extremely important to note that freedom of speech and freedom of expression more generally play a special moral role as humanity-realizing needs or higher-level basic human needs, in that when they are properly practiced, they articulate and disseminate our core dignitarian commitments and humanistic values. But if, in pursuit of this dignitarian aim, someone's digitally-expressed or digitally-stored free speech or other form of self-expression merely offends another human real person and/or draws the ire of those who administer or govern coercive authoritarian social institutions, then this does not entail a morally impermissible design, creation, and/or use of digital technology. Our merely offending someone and/or annoying administrators or governments, is not the same as our oppressing others, whether individually or systemically. The relevant distinction here is between (i) any human real person's being treated in such a way that this mode of treatment fails to have sufficient respect for their dignity or even outright violates their dignity, which is their being oppressed, and (ii) any human real person's merely being greatly annoyed by something that someone else says or writes, or by some other human real person's attitudes or beliefs, even though it does not involve actual oppression, which is their being offended. Being oppressed, therefore, is of real moral significance, whereas being offended is morally insignificant, even if it is of real psychological significance for the human real person who is offended by someone else's speech, attitudes, or beliefs.

5.5 What Privacy Is, and Why Invasions of Digital Privacy Are Morally Impermissible, Whereas Consensual Entrances into Digital Privacy Are Either Morally Permissible or Even Obligatory

To be sure, privacy is not the be-all and the end-all of DDE; nevertheless it *is* important, especially in view of the rise and widespread use of digital technology for legally-permitted ubiquitous data collection—for example, as applied to distance workers during the COVID-19 pandemic (see also Laker et al., 2020).

Now, privacy is the sphere bounded by the living animal body and by the entire life of an individual human real person, including all data or information about that human real person that uniquely identifies them, and especially all digitally-stored data or information that uniquely identifies them; for example, digitally-stored medical data or information, digitally-stored biophysical data or information more generally, digitally-stored photographs or videos of that person,

digitally-stored financial data or information, and digitally-stored social data or information more generally, especially including that person's use of digital technology for surfing the internet and using social media. In view of our moral obligations with respect to all human real persons, no one can morally permissibly enter into the privacy of another human real person without their explicit or implicit rational consent, and in particular, no one can morally permissibly enter into the digital privacy of another person without their explicit or implicit rational consent. Or in other words, rational consent is what freely opens the door into the sphere of privacy, whether this is non-digital or digital. The occurrence of such an entrance without rational consent—an uninvited foot in the door, slipping inside while no one is watching, or breaking down the door—is what I'll call *an invasion of privacy*, and in cases of digitally-stored data or information, *an invasion of digital privacy.* Invasions of digital privacy are violations of sufficient respect for human dignity, and therefore morally impermissible.

It's also extremely important to note that explicit or implicit rational consent can be given for the use of uniquely identifying data or information. Therefore, not every entrance into privacy is morally impermissible. Indeed, a great many entrances into privacy are not only consistent with but even morally obligatory in order to secure sufficient respect for human dignity, and fully involve actual or implicit rational consent: hence I'll call them *consensual entrances into privacy.* In particular, there can be consensual entrances into digital privacy. For example, if a doctor uses digitally-stored uniquely identifying medical data or information about a certain higher-level or Kantian human real person, i. e., a certain moral agent, a use that has been actually or implicitly rationally consented to by that moral agent, in order to save that moral agent's life, then this entrance into digital privacy sufficiently respects that moral agent's dignity, and at the very least it is morally permissible, and perhaps even (say, in view of the doctor's Hippocratic Oath) morally obligatory.

What is the criterion of implicit rational consent? The basic idea is that if any higher-level or Kantian human real person, i. e., any moral agent, were, by means of a thought-experiment, placed behind what Rawls famously called a "veil of ignorance" (Rawls, 1971), which procedurally screens out all uniquely self-identifying personal identity details from that moral agent's own cognitive and practical point of view, and temporarily ensures a suitable reflective disinterestedness and distance from their actual "human, all too human" condition, then they would agree to that treatment. The notion of implicit rational consent is important in cases for which the real-world moral context is so "messy" that the moral agent has little or no opportunity to reflect and make a well-considered judgment; for which the moral agent has, at that time, insufficient knowledge of digital technology; for which there is good reason to believe that the moral agent is not psycho-

logically competent in that particular context; for which there is good reason to believe that the moral agent is being coerced in that particular context; or for which there is good reason to believe that the moral agent is under some or another serious cognitive illusion in that particular context. Since a great many real-world moral contexts are such that explicit rational consent or its refusal is simply out of the question, the notion of implicit rational consent or its refusal plays an essential role in digital privacy.

Now applying this set of ideas to legally-permitted ubiquitous data collection, this is clearly an invasion of digital privacy—it's clearly one thing to evaluate workers' performances, via all-things-considered judgments, but sharply another to spy on them—and therefore morally impermissible, unless the workers explicitly or implicitly rationally consent to this treatment. And it would also be morally impermissible coercion on the part of bosses and management to require workers to agree to this ubiquitous data collection as a necessary condition of accepting job offers or of keeping their jobs. So, yes it is "an Orwellian Big Brother arrangement" (Laker et al, 2020), even if it is legal. And this in turn raises one final important general issue in the area of digital ethics, namely, a basic distinction between dignitarian morality and legality.

5.6 Dignitarian Morality *Versus* Legality, and DDE

Morality, as I've defined it for the purposes of this book, is the attempt to guide human conduct by rationally formulating and following principles or rules that reflect our basic personal and social commitments and our leading ideals and values. But by contrast, legality concerns only what's deemed right or wrong by social institutions (for example, business corporations, universities, governments, judicial systems, etc.) that enforce their judgments, policies, rules, and laws by means of coercion.

But, contrary to an influential trend in moral philosophy, jurisprudence, and political philosophy since the late 1970s (see, e.g., French, 1979), human social institutions (say, business corporations, judicial systems, the police, governments, etc.) are not human real persons, as my definition of human real personhood in section 3.4 above clearly shows, and therefore no matter how coercively powerful and economically profitable they might be, social institutions have no human dignity, i.e., no fundamental moral status or moral value, over and above the human dignity of the individual real human persons who belong to that social institution. And this point holds necessarily, even if so-called "corporate persons" have actually had a conventional legal status that has not only consistently belonged to, but

also in fact has seriously muddled and tainted, the history of moral, legal, and political thinking about persons, for example in the USA (see, e. g., McWhorter, 2017).

To be sure, social institutions themselves are constituted by human real persons with dignity, and therefore social institutions do indeed take on a certain derivative moral status and moral value. But the dignity of human real persons inherently overrides any such social-institutional moral status and moral value. Moreover, and in view of its derivative moral status and value, just because a social institution commands that something is legally right or wrong, it does not thereby follow that it is morally right (good, virtuous, etc.) or wrong (bad, vicious, etc.). For example, as per section 5.5 above, even though digital surveillance of distance workers by business corporations is legally right, nevertheless it is obviously morally wrong. To take another contemporary example, it's plausibly arguable that the social institution of crime-and-punishment in the contemporary USA, especially including the police and prisons, not only provides a morally scandalous example of rationally unjustifiable "systemic racism," and/or "mass incarceration," but also is inherently and systematically authoritarian, coercive, and in violation of human dignity, for anyone who is deemed to be in violation of the laws of the system: hence, while obviously as a system it is legally right, nevertheless equally obviously as a system it's morally wrong (Davis, 2003; Vitali, 2017). Conversely, something can be morally right—say, treating all human real persons with sufficient respect for their dignity—but also legally wrong: for example, and again using an example from the USA, under the system of slavery in the USA prior to The Emancipation Proclamation issued by Lincoln in 1863, treating Black people who were slaves as equal human persons was in fact illegal.

Corresponding to the crucial distinction between morality and legality, is the equally crucial distinction between deep moral or non-moral responsibility and legal responsibility, aka legal accountability. More precisely, (i) deep moral or non-moral moral responsibility belongs fundamentally to higher-level/Kantian human real persons, i. e., moral agents (as opposed to groups of human real persons or to social institutions), by virtue of their choices and/or actions, and in that sense is essentially first-personal, (ii) a moral agent is deeply morally or non-morally responsible for any choice or intention and/or action that flows from that individual moral agent themselves, and (iii) by virtue of the fact that a given choice or intention and/or action flows from that moral agent themselves, then the moral value of that choice and/or action, especially including the moral value of some or all of the consequences of that choice and/or action, also directly attaches to that moral agent themselves. By sharp contrast, legal responsibility or accountability concerns only what people can be deemed and held liable for (accused of, blamed for, punished for, etc.) or what social institutions can be deemed and held liable for, by other people or by other social institutions who/that enforce

their judgments, policies, rules, and laws by means of coercion, hence it is essentially second-personal or third-personal. Obviously, then, a moral agent can be morally responsible for something without also being legally responsible/accountable for it, and also be legally responsible/accountable for something without also being morally responsible for it. So again, the distinction between the fundamental moral status and value of all human real persons (especially including higher-level or Kantian human real persons, i. e., moral agents) and the at-best derivative moral status and value of social institutions is crucial to ethics and morality, not to mention social life and politics.

How does all this apply to DDE? My proposal is simply that legal judgments and legal accountability, as applied to the design, creation, and/or use of digital technology, ought to conform as closely as humanly possible to the basic moral principles of broadly Kantian nonideal dignitarian moral theory. In short, I'm asserting that legality with respect to the design, creation, and/or use of digital technology should be essentially grounded on broadly Kantian nonideal dignitarian morality with respect to the design, creation, and/or use of digital technology, no matter what the consequences might be for business corporations, judicial systems, the police, governments, and so on.

6 How Digital Technology Can Systematically Shape Our Lives in Bad, False, and Wrong Ways, and How Radically Enlightened Dignitarians Can and Should Push Back Against It

> Our age is the genuine age of criticism, to which everything must submit. Religion through its **holiness**, and **legislation** through its **majesty** [and, in the 21st century, **the military-industrial-digital complex that controls the world**, through its **creation and all-pervasive dissemination of digital technology**, its **ideological hegemony**, and its **coercive authoritarianism**], commonly seek to exempt themselves from it. But in this way they excite a just suspicion against themselves, and cannot lay claim to that unfeigned respect that reason grants only to that which has been able to withstand its free and public examination. (*CPR* Axi n., bold-faced text outside the square-bracketted material in the original; square-bracketted material added)

6.1 Introduction

In this chapter, deploying DDE and DCI, I want to apply a broadly Kantian nonideal dignitarian moral-theoretic interpretation of enlightenment that I call *radical enlightenment*, to the design, creation, and use of digital technology in our contemporary sociopolitical world-situation, and try to answer this burning question:

> As radically enlightened people confronted by the design, creation, and use of digital technology by technocratic global corporate capitalism and the military-industrial-digital complex, and in order to push back against cybernetic oppression up to and including cybernetic totalitarianism, what should we dare to think, feel, and do?

The answer that's proposed by DDE and DCI is what I call *cybernetic radical enlightenment*, which in turn is necessarily equivalent to what I call *dignitarian neo-Luddism with respect to digital technology.*[29]

29 See chapter 1, n. 3 for important caveats; see also chapter 7 for an explicit definition of "dignitarian neo-Luddism with respect to digital technology."

https://doi.org/10.1515/9783111260525-009

6.2 The Life-Shaping Thesis

As rational human animals, we design, create, and use social institutions in order to satisfy human needs (ideally, true human needs, as per section 5.3 above, but often enough, false, or misguided and self-stultifying, human needs), by communicating and engaging with other people in an open-ended plurality of group enterprises, practices, and projects, according to shared normative rules, for better or worse; a principal example of this is human language. Also as rational human animals, we design, create, and use cognitive tools, principally including thought-shapers (Hanna and Paans, 2021), propositional mental representations, chains of reasoning, and theories, in order to refer to the world, accurately or inaccurately, and/or to interpret the world, truly or falsely. And we also design, create, and use physical tools and machines—i.e., technology, nowadays principally including digital technology—in order to change the world, again for better or worse. But as a direct result of all this rational human design, creation, and use, undertaken by us in order to satisfy our (true or false) human needs, refer to the world, interpret the world, and change the world, nevertheless, at the same time, more or less unintentionally and unselfconsciously, our social institutions, our cognitive tools, and our technology also all inevitably change us, for better or worse. I'll call this *The Life-Shaping Thesis* (see also Maiese et al, 2023). The Life-Shaping Thesis has two parts.

The first part is a comprehensive fusion of philosophy of mind and cognition and sociopolitical theory, called *the mind-body politic* (Maiese and Hanna, 2019) which has seven basic sub-theses.

1. Human minds are necessarily and completely embodied, and identical to the complex dynamic intentional-action-guiding structures of suitably complex living organisms, i.e., human animals (the essential embodiment thesis) (Hanna and Maiese, 2009; Hanna, 2011a).
2. As essentially embodied and inherently dynamic, human minds are also inherently enactive and environmentally embedded (the enactivity-and-embeddedness thesis) (Hanna and Maiese, 2009; Newen, De Bruin, and Gallagher, 2018).
3. Human animals are, necessarily, social animals (the human sociality thesis) (Maiese and Hanna, 2019: ch. 1; Hanna, 2021b).
4. Social institutions partially determine and form our essentially embodied minds, and thereby substantially affect and guide our human lives (the mind-shaping thesis) (Maiese and Hanna, 2019: ch. 2).
5. There is a fundamental distinction between (5.1) destructive, deforming social institutions that frustrate and warp true human needs, and (5.2) constructive, enabling social institutions that satisfy and sustain true human

needs (the two-kinds-of-social-institutions thesis) (Maiese and Hanna, 2019: esp. chs. 2–3 and 6–8).

6. Enacting salient changes in the structure and complex dynamics of a social institution produces corresponding salient changes in the structure and complex dynamics of the essentially embodied minds of the participants, for better or worse (the enactive-transformative thesis) (Maiese and Hanna, 2019: esp. chs. 2–3 and 6–8).
7. Although destructive, deforming social institutions shape human minds and human lives in an inherently bad/oppressive, unhealthy, and enslaving/heteronomous way, nevertheless it is also possible to devolve such institutions and also simultaneously to create constructive, enabling social institutions that are inherently good/non-oppressive, healthy, and emancipatory/autonomous (the social devolution-social creation thesis) (Maiese and Hanna, 2019: esp. chs. 2–3 and 6–8; Hanna, 2018c: esp. parts 2–3).

And the second part of The Life-Shaping Thesis is a dual application of the mind-body politic, together with the mechanistic worldview vs. neo-organicist worldview distinction, to a comprehensive philosophy and psychology of human thinking, the Theory of Thought-Shapers, aka TTS (Hanna and Paans, 2021). TTS applies the categorical distinction between (i) mechanical (i.e., computable/recursive, entropic, and deterministic or indeterministic) systems, and (ii) organic (i.e., uncomputable/processual, negentropic, purposive, and self-organizing) systems, to fundamental issues in the philosophy of mind and cognition, with comprehensive application to the nature of human thinking as such. By *thought-shapers*, I mean any or all of the following kinds of essentially non-conceptual, non-discursive, non-propositional mental representations: allegories, analogies, blueprints, catechisms, diagrams, displays, icons, images, lay-outs, metaphors, mnemonics, models, outlines, parables, pictures, scenarios, schemata, sketches, spreadsheets, stereotypes, symbols, tableaux, and templates.[30] In turn, TTS says that all human thinking is really possible only insofar as it's partially causally determined, formed, and normatively guided by either (i) mechanical, constrictive thought-shapers in a bad, false, and wrong way, or (ii) organic, generative thought-shapers in a good, true, and right way. Now, since human language is a core social institution, and since all human thinking proceeds by means of human language, TTS falls directly under the mind-body politic; and if the mind-body politic is true, then both mind-shaping inside social institutions and also thought-shaping in individuals and groups, alike, are forms of life-shaping, for better or worse.

30 This is intended as an open-ended, partial list only.

In view of The Life-Shaping Thesis, what I want to argue in the next four sections is that digital technology can systematically shape our lives in bad, false, and wrong ways.

6.3 John von Neumann and Richard Feynman

As Herman Goldstine points out in *The Computer from Pascal to von Neumann*, after Turing, John von Neumann is arguably the most important figure in the theoretical foundations of computer science (Goldstine, 1980: esp. pp. 167–178).[31] Von Neumann was also a key figure in the Manhattan Project, by means of which the USA designed, created, and disastrously deployed the atomic bomb against Japan—twice—in 1945. This is what von Neumann said about the Project during that time, the crucial phrases of which are the highlighted ones:

> What we are creating now is a monster whose influence is going to change history, provided there is any history left, yet it would be impossible not to see it through, not only for the military reasons, but *it would also be unethical from the point of view of the scientists not to do what they know is feasible, no matter what terrible consequences it may have.* (von Neumann, as reported by his wife and quoted in Dyson, 2012: p. 62, italics added)

In other words, von Neumann is saying that whatever scientists can do, they should do, no matter what the consequences, and no matter how intrinsically bad and wrong these things might be. If von Neumann had been working for the Nazis, then the same principle would have held. So, von Neumann was the *Dr Frankenscience* of digital technology and the atomic bomb. In turn and in a closely related way, Richard Feynman, who was a young and rising physicist during the Manhattan Project era, said this about von Neumann:

> [v]on Neumann gave me an interesting idea; that you don't have to be responsible for the world you're in. So I have developed a very powerful sense of irresponsibility as a result of [v]on Neumann's advice. It's made me a very happy man ever since. (As quoted in Dyson, 2012: p. 62)

More generally, then, von Neumann was a textbook case of someone who, when he applied mechanical, recursive thought-shapers to mechanical recursive domains of content, was absolutely brilliant. But when he applied the very same thought-shapers to organic, non-mechanical domains of content, and especially to human, moral, and sociopolitical domains of content, his thinking became disas-

31 See also McKeever, 2021.

trously mechanical and constrictive, and monstrously bad, false, morally wrong, and directly contrary to the universal dignitarian obligation always to treat everyone with sufficient respect for their human dignity. So in a behavioral and functional sense, von Neumann turned himself into a walking, talking, researching, and publishing incarnation of bad, false, and wrong life-shaping digital technology: and he also spread these monstrously bad, false, and wrong mechanical, constrictive thoughts to his younger, impressionable colleagues, like Feynman.

6.4 Nuclear Weapons Design and Virtual Reality

In a critically incisive study of the use of digital simulation, Sherry Turkle writes:

> In the years after the 1992 ban, newcomers to the field of nuclear weapons design would see explosions only on computer screens and in virtual reality chambers. At Lawrence Livermore and Los Alamos National Laboratories, some of the most powerful computer systems in the world are used to simulate nuclear explosions. Until recently, these simulations took place in two dimensions; now, simulations are moving into three dimensions. In a virtual reality chamber known as a CAVE, one stands "inside" a nuclear explosion wearing 3D goggles, in order to observe it, one is tempted to say, "peacefully." My story of simulation began with the Athena project centered in a garden, a glass atrium with a ficus tree; it ends in a CAVE, a self contained virtual reality. The CAVE simulation is there to "demo" an explosion; those who work there become accustomed to experiencing in the virtual what could never be survived in the real.
>
> When nuclear testing moved underground, it became easier for weapons designers to distance themselves from the potential consequences of their art. Hidden, the bomb became more abstract. But even underground testing left craters and seismic convulsions. It scarred the landscape. Now, with explosions taking place on hard drives and in virtual reality chambers, how much harder will it be for weapons scientists to confront the destructive power of their work and its ethical implications? One weapons designer at Livermore laments that he has only once experienced "physical verification" after a nuclear test. He had "paced off the crater" produced by the blast. It changed him forever. His younger colleagues will not have that.
>
> This senior scientist is concerned about the moral effects of moving nuclear weapons research to virtual space, but he and his colleagues are also troubled about the effects of virtuality on their science itself. They argue that "physical intuition is a skill you want to keep" and worry that the enthusiastic reactions of young designers to new, Flashy virtual reality demonstrations are naïve. One says: "The young designers look at anything new and they say, 'This is so much better than what we had before. We can throw out everything we did before!'" Senior scientists at the national laboratories describe young designers immersed in simulation as "drunk drivers."
>
> Within simulation, the happily inebriated show less judgment but think they are doing fine. Dr. Adam Luft, a senior weapons designer at Los Alamos, shows sympathy for the young de-

> signers: the new rules compel them to fly blindly. They cannot test their weapons because they must work in the virtual and they are given computer systems whose underlying programs are hard to access. Luft himself feels confident only if he is able to access underlying code. He is frustrated by the increasingly opaque simulations of his work environment. When something goes wrong in a simulation, he wants to "dig in" and test aspects of the system against others. Only a transparent system "lets [me] wander around the guts of [a] simulation." He is wary of making any change to a weapon without personally writing its code. Luft worries that when scientists no longer understand the inner workings of their tools, they have lost the basis for trust in their scientific findings, a concern that mirrors those of MIT designers and scientists of twenty years before. (Turkle, 2009: pp. 72–73)

Interestingly, the senior scientists that Turkle talked to, said that the young nuclear weapons designers lacked "physical intuition" and that their simulations lacked "transparency." But if The Life-Shaping Thesis is correct, then that's only part of the problem and not even the core of the problem. For armies have been slaughtering each other face-to-face for millennia, even though they usually have more or less unobstructed "physical intuition" and "transparency." The core of the problem with these nuclear scientists is that the digital technology they're using is shaping their thinking in mechanical, constrictive, disastrously, and indeed monstrously bad, false, and wrong ways.

6.5 US Air Force Killer-Drone Operators

Now let's consider confessional whistle-blowing reports by former USAF killer-drone operators, for example, Brandon Bryant:

> Brandon Wayne Bryant … is an American whistleblower. From 2006 to 2011 he was a camera operator (sensor operator) of unmanned drones of the United States Air Force. His job was targeted killing. On December 10, 2012 the German magazine *Der Spiegel* reported about Bryant, his former work and his post traumatic stress disorder. He gave his testimony to an expert council of the United Nations and to the German Parliamentary Committee investigating the NSA spying scandal. Jesselyn Radack is his lawyer.
>
> The smoke clears, and there's pieces of the two guys around the crater. And there's this guy over here, and he's missing his right leg above his knee. He's holding it, and he's rolling around, and the blood is squirting out of his leg … It took him a long time to die. I just watched him. (Airman First Class Brandon Bryant in GQ)
>
> …. In May 2016 the Norwegian documentary filmer Tonje Hessen Schei produced the documentary *Drone—This Is No Game!* Drone pilots Brandon Bryant and Michael Haas are shown as examples of how the CIA recruits video gamers and trains them to kill by remote control. (Wikipedia, 2025f)

Thus the USAF recruited former video-gamers, and trained them/mind-controlled them by getting them to focus on killing the target-images on the screens, as if they were super-powered video games, with schizoid-like compartmentalization, and almost complete emotional alienation from real-world impact and the people they were actually killing. And of course, standing behind that digital technology, is the destructive, deforming social institution of the US military. So if you're life-shaped by digital technology in general and by the US military digital-technological training regime in particular, into thinking about other (and especially non-US) human beings as "moist robots" (Dennett, as quoted in Schuessler, 2013) as mere mechanical things and not living people with dignity, then of course it's all-too-easy to shift into "von Neumann-mode" and convert those shaped thoughts into "physical intuition" and "transparent" interaction with the real world, and thereby act like a real-world Terminator.

6.6 The New Wheel of Ixion: Work, Sleep, and the Mechanization of Everyday Life

In Greek mythology, the "human, all too human" King Ixion was bound by the all-powerful Zeus, King of the Gods on Mount Olympus, to a burning, fiery wheel for eternity, as punishment for abusing Zeus's hospitality and seducing the *über*-God's wife. The Wheel of Ixion initially rolled across the sky, but was eventually moved below to Tartarus, the better to darken and increase Ixion's suffering, just like the exceptionally cruel and more famous punishments also meted out to Sisyphus and Tantalus. Arthur Schopenhauer was a pessimistic philosophical genius of the early-to mid-19th century, who, in his brilliant essay, "On the Sufferings of the World" (Schopenhauer, 1851/2004), used Ixion's Wheel as a vivid metaphor for the endless cycle of burning desire, its frustration or merely temporary satiation, followed immediately by renewed desire, pain, and suffering, on and on and on relentlessly, and finally death, that's built into the "human, all too human," condition and the "will-to-live" (*Wille zum Leben*) alike. Now I'm going to describe a 21st-century update of these ancient Greek and Schopenhauerian visions of our "human, all too human" condition, that I call The New Wheel of Ixion.

If you are like most adult people alive today (I mean all those adults who have jobs: for simplicity's sake I am leaving aside the important but also complicating issue of temporary or permanent joblessness or unemployment), then (i) you work at a salaried job for more or less eight continuous hours a day, five days a week (similarly, and again for simplicity's sake, I am leaving aside the equally important but complicating issue of part-time jobs and overtime work) and (ii) you sleep more or less eight continuous hours every single day (ditto, for sleep disor-

ders), and, because your work occurs at some distance from your home, then (iii) you also spend at least another two hours getting ready to work, traveling to-and-from work, getting ready to sleep, and or slowly waking-up. Or, as during the COVID-19 pandemic and beyond, if you're distance-working at home and not traveling, then you spend another two hours per day working, probably unpaid. Of course, there are usually weekends off, or at least two discontinuous days off, every week, and a few weeks off for holidays, every year; but the unending, unrelenting work-and-sleep pattern continues throughout. So the stunning fact is that most human adults spend up to or even more than 2/3 of their lives continuously working, continuously sleeping, or going through the supplementary rituals associated with work and sleep.

By *jobwork*, aka *labor*, I mean human work that consists in "the activities necessary to society's survival" (Kranzberg, 1981: p. 932). But, sadly, most jobwork/labor is soul-destroying. Why do I say that? I say it, because it's a plain and simple fact that most jobs are either (i) bullshit, that is, absurdly pointless and also unproductive and useless for anyone other than the job-holder—even if well-paying (Graeber, 2013, 2018), (ii) fairly or very boring, or (iii) even worse: badly-paid, dangerous, demeaning, or otherwise exploitative or oppressive. So, more generally, I'll call any job that is either bullshit, fairly or very boring, badly-paid, dangerous, demeaning, or otherwise exploitative or oppressive, a *soul-destroying job*.

In *Kant, Agnosticism, and Anarchism*, I defined human work as follows:

> any form of creative, productive, or otherwise energy-expending rational human agency or performance (roughly, intentionally changing or moving oneself or other things, in the natural or social worlds). (Hanna, 2018c: p. 122)

And in that same book I also said that there are two basic kinds of work, namely, (i) what I call *jobwork*, aka labor, and (ii) what I call *lifework*, and that

> [l]ifework … is … creative, meaningful activity (aka a project), or a series of such activities (aka projects), pursued as a full-time, or almost full-time, lifetime calling. Simply put, lifework is whatever you would do for the rest of your life if you were freed from financial worries. (Hanna, 2018c: p. 122)

But in order to be more complete and precise, I'll now define human work like this, with the changed bits italicized:

> any form of creative, productive, or otherwise energy-expending rational human agency or performance (roughly, *either* intentionally changing or moving oneself, *other people*, or other things, in the natural or social worlds, *or intentionally keeping oneself, other people, or things in their current conditions*

or positions, or preventing their change or movement, in the natural or social worlds).

Correspondingly, I'll also now categorize and sub-categorize the basic kinds of human work as follows: (i) jobwork, aka labor: human work that consists in "the activities necessary to society's survival" (Kranzberg, 1981: p. 932), including (ia) technocratic global corporate capitalist jobwork/labor, in exchange for money or some equivalent form of payment (oppressive, inconsistent with lifework), and (ib) post-capitalist jobwork/labor,[32] with payment (non-oppressive, consistent with lifework), (ii) slavework: coerced, compelled work, with or without payment (oppressive, inconsistent with lifework), (iii) domestic work: human work for creating, maintaining, or sustaining a household, including looking after partners, children, or other family members, with or without payment, including (iiia) oppressive domestic work, and (iiib) non-oppressive domestic work, (iv) basic-needs work: human work for providing what is necessary for the satisfaction of basic true human needs, including food, shelter, clothing, healthcare, other public services like transportation/ thoroughfares, etc., with or without payment, including (iva) oppressive basic-needs work, and (ivb) non-oppressive basic-needs work, and (v) lifework: creative, meaningful activity (aka a project), or a series of such activities (aka projects), of any kind, done for its or their own sake, with or without payment, pursued as a full-time, or almost full-time, lifetime calling; what you would freely choose to do for your entire working life, if liberated from financial worries, worries about social status, slavery, domestic work worries, and basic-needs work worries.

Notice particularly in the list of categories of work, that on the one side, (ia) and (ii), and on the other side, (v), are contraries, and therefore mutually inconsistent. Recognizing this antithesis then enables the further recognition that there's a

32 It's essential to contrast technocratic global corporate capitalism with a radical alternative I call *dignitarian post-capitalism* (Hanna, 2023b), in which there's (i) modest individual ownership of private property, sufficient for satisfying and sustaining people's true human needs and individual tastes, (ii) non-commodifying, non-exploitative, non-wage-enslaving (for example, under a larger system of truly generous universal basic income and universal basic jobs [Hanna, 2018c: sections 3.3–3.4]), eco-sensitive/sustainable, small-scale business enterprises for the production of goods and the provision of services, sufficient for satisfying and sustaining people's true human needs and individual tastes, (iii) modest individual profit-making accumulation of wealth, sufficient for satisfying and sustaining people's true human needs and individual tastes, (iv) modest, non-commodifying, non-exploitative, non-wage-enslaving, collective profit-making and collective profit-sharing enterprises (aka cooperatives), sufficient for satisfying and sustaining workers' true human needs and individual tastes, and (v) a robust, worldwide commons system for interpersonal, public, and cosmopolitan sharing (Ostrom, 1990; P2P, 2025).

profound sociohistorical and political connection between slavework and technocratic global corporate capitalist jobwork/labor: the social institution of chattel slavery is when some people own other people as private property, and the slave-owners have coercive authoritarian control over the lives of those people by forcing them to work solely for their owners' self-interest; so chattel slavery was the historical precursor or prototype of technocratic global corporate capitalist jobwork/labor.

Correspondingly, let's now turn from this brief critical analysis of *how we work* during the era of technocratic global corporate capitalism and digital technology, to an equally brief critical analysis of *how we sleep* during this era. For five days a week, during your more or less eight continuous hours of sleep, exhausted by your jobwork and by your two hours of associated jobwork rituals, together with the rest of your life squeezed into the other more or less six continuous hours, you do nothing but lie there, either dreamlessly or merely passively dreaming, meaninglessly alive, listlessly rolling over from time to time. And the same sleep pattern recurs over the weekend, precisely in order that you can laboriously resume it on Blue Monday, *ad infinitum*.

In fact, most ordinary people in the Middle Ages and other pre-industrial societies naturally engaged in multiphasic, polyphasic, or segmented restfulness or sleep, usually in a biphasic pattern consisting of (i) roughly four hours of sleep (which was often called "the first sleep"), (ii) roughly two hours of contemplative or meditative, or in any case, restful wakefulness in the middle of the night (for lack of a better term, and with Andrei Tarkovsky's exceptionally eerie and hypnagogic 1979 post-apocalyptic science fiction film *Stalker* in mind, let's call this "The Zone") and (iii) another roughly three or four hours of sleep (which was often called "the second sleep") (Ekirch, 2001, 2005, 2016; Yetish et al., 2015). Correspondingly, in the mid-17th century, Blaise Pascal was convinced that sleep is inherently organic and vital, not mechanical and deathlike:

> Sleep, you say, is the image of death; for my part I say that it is the image of life. (Pascal, 1995: p. 333)

I'll call the act and process of sleeping according to this Pascalian organicist conception of sleep, *lifesleep*. Of course, human patterns and rhythms of sleeping changed radically with the gradual advance and eventual domination of capitalist industrialization. Indeed, endorsing but also extending the main line of argument in Jonathan Crary's excellent book, *24/7: Late Capitalism and the Ends of Sleep* (Crary, 2013), I'm claiming that our patterns and rhythms of sleeping have been social-institutionally life-shaped (Maiese and Hanna, 2019; Maiese et al., 2023; see also section 5.2) by technocratic global corporate capitalism and digital technology

since 1950, as these have gradually turned Eisenhower's military-industrial complex into the military-industrial-digital complex (see chapter 1, n. 2). And insofar as our natural, organic sleep patterns have been social-institutionally mind-shaped and life-shaped in this way, that seventy-plus years regime of life-shaping has been systematically destructive and deforming for us (Maiese and Hanna, 2019: chs. 2–3). In short, the military-industrial-digital complex has turned our organic lifesleep into mechanical zombie-sleep.

So what I'm calling *The New Wheel of Ixion* is the stunning real-world fact that most adult human beings are permanently locked into an utterly commodified, mechanized, digitally technologized, and more generally robotic routine consisting of (i) more or less eight or more continuous hours of working at a soul-destroying job five days a week, much of it spent either sitting in front of a computer screen in an office or while distance-working at home, or messing around on their smart phone in either location, (ii) another one to two meaningless hours of associated jobwork rituals supplementary to those, for five days a week, (iii) weekends spent trying to catch up with the rest of their life, but often enough including overtime work at the office or online, or worrying about work, and (iv) more or less eight continuous hours of passive, meaningless sleep seven days a week. And thereby, most of us spend up to, or even more than, 2/3 of our lives either soul-destroyingly or passively and meaninglessly, until so-called "retirement," or death, whichever comes first. No wonder, then, that many or most contemporary professional academic philosophers, formal scientists, natural scientists, cognitive neuroscientists, and social scientists, all working together under the big tent of the strong AI program, billionaire technocrats, military leaders, and politicians, all of them dogmatically gripped by the hegemonic ideologies of scientism and mechanism, endlessly tell us that we're nothing but decision-theoretic biological machines—"biochemical puppets" (Harris, 2012), "moist robots" (Dennett, as quoted in [Schuessler, 2013]), or "survival machines" (Dawkins, 2006). Now, by a *placebo*, I mean anything X, the mere (usually false) belief in which causes people to be morally better, or better off, than they would have been without having that belief in X; and by a *nocebo*, I mean anything Y, the mere (usually false) belief in which causes people to be morally worse, or worse off, than they would have been without having that belief in Y (see also Bregman, 2020: pp. 8–9, 17, 37, 134, 228, 249, 258, 270, and 395). Then even if, as I've shown in chapter 2, metaphysically it cannot be the case that we're nothing but squishy Turing machines, nevertheless this false belief operates as an exceptionally powerful nocebo for shaping our lives in just those bad, false, and wrong ways that best serve the ends of the military-industrial-digital complex.

6.7 From Classical Enlightenment to Cybernetic Radical Enlightenment

To coin a question, *what is enlightenment?* In his equally famous and notorious, but in any case historically and philosophically seminal, same-named essay, Kant says this:

> *Enlightenment is the human being's emergence from his own self-incurred immaturity. Immaturity* is the inability to make use of one's own understanding without direction from another. This immaturity is *self-incurred* when its cause lies not in lack of understanding but in lack of resolution and courage to use it without direction from another. *Sapere aude!* Have the courage to use your *own* understanding! is thus the motto of Enlightenment. (*WiE* 8: 35, italics in the original)

Following Kant's lead and elaborating a little, I'll say that the classical concept of ENLIGHTENMENT, as such, says that (i) because we are, by virtue of a unified set of innately-specified cognitive, practical, and affective/emotional capacities, autonomous rational human animal moral agents inherently possessing dignity, that is, rational human minded animals and human real persons, but (ii) because, tragically, against the backdrop of various more or less brute, goading material conditions of physical nature, human history, and social-institutional or political culture, we have also freely—even if unreflectively and self-deceivingly—put ourselves into a longstanding condition of cognitive, practical, and affective/emotional passivity, robotic subservience, mind control, and mental slavery, and thus into a longstanding state of self-incurred moral and intellectual immaturity, (iii) therefore, in order finally to advance beyond this tragic immature condition and to satisfy the categorically normative demands of our rational human nature as persons, we ought to dare to use our own understanding and think for ourselves, or as the classical slogan has it, *Sapere aude!*

But unfortunately, and fatefully, whether intentionally or not, Kant's seminal essay is highly ambiguously written, in such a way as to permit two sharply different readings of the classical concept of enlightenment, depending on whether one interprets it, as most casual readers, scholars, and Kantian or non-Kantian philosophers do, in the light of (i) Kant's neo-Hobbesian liberal Statist political philosophy in *The Doctrine of Right*, or instead, as a few "Left Kantians" (Hanna, 2017b, 2017c) do, in the light of (ii) Kant's uncompromising non-egoistic, non-consequentialist, autonomy-driven, dignitarian ethics in the *Groundwork* and *Critique of Practical Reason*, his post-Statist, spiritually-inspired moral cosmopolitanism in *Religion Within the Boundaries of Mere Reason*, and his defense of the absolute autonomy of philosophy in "The Conflict of the Faculties."

According to the first interpretation, which I call Enlightenment Lite (EL), you ought to "argue as much as you like about whatever you like, but obey!" Correspondingly, EL is committed to an essentially instrumental, empiricist conception of cognitive and practical rationality, an essentially deterministic or at least compatibilist metaphysics of free will and autonomy, an essentially egoistic utilitarian conception of ethics, an essentially individualist conception of social life, and an essentially intellectualist or dualist conception of the nature of the human mind. But on the contrary, according to the second interpretation, which I call *Radical Enlightenment* (RE),[33] or *Heavy-Duty Enlightenment*, you ought to dare to think and care (including desires, emotions, and feelings) and act for yourself, and in so thinking, so caring, and so doing, thereby exit the State in order to create and sustain a cosmopolitan moral community that Kant calls the "ethical community": hence RE is a broadly Kantian, cosmopolitan, dignitarian anarcho-socialism (Hanna, 2016, 2017b). Correspondingly, RE is committed to an essentially non-instrumental, apriorist conception of theoretical cognition and rationality (Hanna, 2006b, 2006c, 2015a), a natural libertarian, source-incompatibilist metaphysics of free will and practical agency (Hanna, 2018b), a robustly dignitarian conception of morality and moral theory (Hanna, 2018c; see also chapters 3–5 above), an essentially embodied and caring-oriented conception of the human mind (Hanna and Maiese, 2009), and an essentially enactive and embedded conception of social-institutional life (Maiese and Hanna, 2019).

Now the ideological allure of the first or EL interpretation is so powerful that you may find it hard to believe that there even is a second or RE interpretation. If so, then I hereby invite you to put the all-too-familiar EL interpretation in abeyance for a very brief moment, and recognize that the RE interpretation practically leaps out of these three juxtaposed texts:

> When nature has unwrapped, from under this hard shell, the seed for which she cares most tenderly, namely the propensity and calling to *think* freely, the latter gradually works back upon the mentality of the people (which thereby gradually becomes capable of *freedom* in acting) and eventually even upon the principles of *government*, which finds it profitable to itself to treat the human

33 In his excellent but also highly controversial *Radical Enlightenment: Philosophy and the Making of Modernity 1650–1750* (Israel, 2001), and its two sequel volumes, Jonathan Israel traces the origins of the very idea of a radical enlightenment project back to Spinoza, pantheism, and metaphysical monism. I certainly agree with Israel that Spinozism is at least *one* important source of the radical enlightenment tradition, but also think that Kant is a more important philosopher than Spinoza in every relevant respect. In any case, Kant's own contribution to the controversy about Spinozism is presented in "What Does It Mean to Orient Oneself in Thinking?" (*OT*).

being, *who is now more than a machine*, in keeping with his dignity. (*WiE* 8: 41–42, italics in the original)

A juridico-civil (political) *state* is the relation of human beings to each other inasmuch as they stand jointly under *public juridical laws* (which are all coercive laws). An *ethico-civil* state is one in which they are united under laws without being coerced, i. e., under *laws of virtue* alone.... In an already existing political community all the political citizens are, as such, still in the *ethical state of nature*, and have the right to remain in it; for it would be a contradiction (in *adjecto)* for the political community to compel its citizens to enter into an ethical community, since the latter entails freedom from coercion in its very concept. Every political community may indeed wish to have available a dominion over minds as well, according to the laws of virtue; for where its means of coercion do not reach, since a human judge cannot penetrate into the depths of other human beings, there the dispositions to virtue would bring about the required result. But woe to the legislator who would want to bring about through coercion a polity directed to ethical ends! For he would thereby not only achieve the very opposite of ethical ends, but also undermine his political ends and render them insecure.—The citizen of the political community therefore remains, so far as the latter's lawgiving authority is concerned, totally free: he may wish to enter with his fellow citizens into an ethical union over and above the political one, or rather remain in a natural state of this sort.... THE HUMAN BEING OUGHT TO LEAVE THE ETHICAL STATE OF NATURE IN ORDER TO BECOME A MEMBER OF AN ETHICAL COMMUNITY. (*Rel* 6: 95–102, italics and capital letters in the original)

When it is a question of the *truth* of a certain teaching to be expounded in public, the teacher cannot appeal to a supreme command nor the pupil pretend that he believed it by order. This can happen only when it is a question of action, and even then the pupil must recognize by a *free* judgment that such a command was really issued and that he is obligated or at least entitled to obey it; otherwise, his acceptance of it would be an empty pretense and a lie. Now the power to judge autonomously—that is, freely according to principles of thought in general)—is called reason. So the philosophical faculty, because it must answer for the truth of its teachings it is to adopt or even allow, must be conceived as free and subject only to laws given by reason, not by the government. (*CF* 7: 27, italics in the original)

In any case, a striking contemporary example of EL is Steven Pinker's *Enlightenment Now* (Pinker, 2018); and, by sharp contrast, I've developed and defended RE in *Kant, Agnosticism, and Anarchism* (Hanna, 2018d).

Very much in the spirit of RE, in "The Conflict of the Faculties," Kant is fully critically aware of the real possibility of mind-control and mental slavery (aka "ideological hegemony") within the university, especially via the faculty of theology; and he correspondingly asserts the absolute autonomy of the faculty of philosophy from the theology faculty, from other university faculties, and from the government. But, as radical as that is, perhaps surprisingly, Kant neglects to consider the equally real possibility of ideological hegemony within faculties of philosophy themselves (Hanna, 2021a: ch. XVIII), nor—perhaps unsurprisingly, as an 18th-century philosopher—does he foresee the real possibility of almost unlimited mind-control and mental slavery via digital technology, not only within States but also across States, worldwide.

When we combine (i) the coercive authoritarianism of all States, especially including all contemporary neoliberal States, with (ii) technocratic global corporate capitalism, and with (iii) globalized digital technology, then we have the military-industrial-digital complex: the global power elite that controls contemporary neoliberal States and their governments, very often more or less covertly and without our being self-consciously critically aware of it—the globalized 21st-century version of the totalitarian State of Alphaville. As I mentioned in section 6.1 above, in this chapter I want to work out an answer to this burning question:

> As radically enlightened people confronted by the design, creation, and use of digital technology by technocratic global corporate capitalism and the military-industrial-digital complex, and in order to resist cybernetic oppression up to and including cybernetic totalitarianism, then what should we dare to think, feel, and do?

And as I also mentioned in section 6.1 above, the answer that's proposed by DDE and DCI is what I call *cybernetic radical enlightenment*, which in turn is necessarily equivalent to dignitarian neo-Luddism with respect to digital technology.[34] I'll present this in three stages.

---

34 Again, see chapter 1, n. 3 above for important caveats and also chapter 7 below for a fully explicit definition of "dignitarian neo-Luddism with respect to digital technology."

6.8 Cybernetic Radical Enlightenment, Stage I: Sociopolitical Critique, Logging Off, and Subverting-Dismantling-&-Replacing

I'll start off with a critical analysis of digital technology in the form of social media, by Benjamin Y. Fong in the American democratic socialist journal, *Jacobin.* Fong writes:

> For the Left, ... social media presents an imminent threat: it attracts people who are natural fodder for socialist politics and then absorbs them in the unthinking narcissism of pseudo-political statement pronouncement, where they enter the negative feedback loop that distances them from the reality of everyday human engagement. Twitter is thus not just a *medium of expression* for the "psychic pathologies" of what Mark Fisher described so well as the "Vampire Castle" [Fisher, 2013]. *It is the Vampire Castle,* doing capitalism's work by further atomizing and distancing people from the kinds of conversations required for real political engagement. The sooner we realize this about social media, the sooner we can get to the work of dismantling it. (Fong, 2018; square-bracketted material added)

Here, in turn, is a four-step rational reconstruction of Fong's argument:

1. Socialism—whether democratic socialism or social anarchism (aka anarcho-socialism, libertarian socialism, etc.)—is fundamentally concerned with respect for universal human dignity; with human freedom of thought, expression, choice, and action; with individual and collective creativity and flourishing; and with the universal satisfaction of true human needs.
2. Internet-based social media may appear to be highly promising and legitimate vehicles for the realization of socialist aims.
3. But in fact, social media are an essential part of the military-industrial-digital complex that not only produces widespread mind-control and mental slavery, but has also enabled a worldwide mental health crisis of social media addiction (Schulson, 2015; Griffiths, 2018; Nguyen, 2018).
4. Therefore, anyone who recognizes the value of the fundamental concerns of socialism should (4i) engage in a sociopolitical critique of social media, (4ii) "log the fuck off" on a regular basis, in order to protect oneself against their bad, false, and wrong life-shaping influence, and (4iii) wholeheartedly individually and collectively commit to subverting and dismantling the entire system of social media and replacing it with something essentially better.

I think that this argument is sound. Moreover, I also think that its conclusion should be generalized so as to apply to all digital technology controlled by technocratic global corporate capitalism and the military-industrial-digital complex, not just social media, therefore all digital technology, including all parts of the inter-

net, that are controlled by (i) technocratic global corporate capitalists and (ii) the governments of contemporary States, especially including their coercive authoritarian enforcement-specialists, the military and the police, insofar as they're driven by the military-industrial-digital complex. The rationale for this generalization is that premises 1, 2, and 3 of the above argument, when conjoined with the moral principles of DCI, apply just as correctly and directly to all digital technology controlled by technocratic global corporate capitalism and the military-industrial-digital complex, as they do to social media in particular. Therefore, by appropriate substitutions into the conclusion of the above argument, step 4, we derive:

> 4*. Therefore, anyone who recognizes the value of the fundamental concerns of DCI should (4i*) engage in a sociopolitical critique of any and all kinds of digital technology designed, created, or used for cybernetic oppression, (4ii*) "log the fuck off" this digital technology on a regular basis, in order to protect oneself against its bad, false, and wrong life-shaping influence, and also (4iii*) wholeheartedly individually and collectively commit to subverting and dismantling any and all kinds of digital technology used for cybernetic oppression and replacing them with something essentially better.

6.9 Cybernetic Radical Enlightenment, Stage II: Civil Disobedience

Now for stage II. In chapter 1, I argued that no matter how powerful, sophisticated, or profitable it is, no digital technology is inevitable, and that if any digital technology violates human dignity, then we should abolish it, refuse it, or at the very least radically transform it into its moral opposite. In this section, I'll work out the basic details of how we can collectively engage in this abolition, refusal, and radical transformation of dignity-violating digital technology.

Before I get down to the nitty-gritty argumentation, however, I want to define some terminology: *Statism*, *coercion*, and *authoritarianism*.

As Kant (*MM* 6: 203–372) and Max Weber have correctly pointed out (Weber, 1994: p. 310), all States possess a territorial monopoly on the (putatively) legitimate control of the means and use of coercion; and as philosophical and political anarchists from the 18th century (William Godwin) to the 19th and early 20th century (Peter Kropotkin) and into the 21st century (e.g., Hanna, 2018c) have also correctly pointed out, States are also inherently authoritarian. By *coercion*, I mean: either (i) using violence (for example, injuring, torturing, or killing) or the threat of violence, in order to manipulate people against their will according to certain predefined purposes of the coercer (primary coercion), or (ii) inflicting appreciable, salient

harm (for example, imprisonment, termination of employment, or large monetary penalties) or deploying the threat of appreciable, salient harm, even if these are not in themselves violent, in order to manipulate people against their will according to certain predefined purposes of the coercer (secondary coercion). So all coercion is a form *of manipulation*, and proceeds by following a variety of strategies that share the same core characteristic: treating people as mere means or mere things. Correspondingly by *authoritarianism*, I mean the doctrine that telling people to obey commands and do things is legitimated merely by virtue of the fact that some people (the purported authorities) have told them to obey those commands or do those things—"it's right just because we say it's right!"—and are also in a position to enforce this by means of coercion, not on any rationally justified or objectively morally defensible grounds. Therefore, authoritarianism and coercion per se are non-synonymous and logically independent, because although all authoritarianism requires coercion, nevertheless the converse is not the case: coercion can occur without authoritarianism—for example, if you're threatened or attacked on the street by some random thug.

The crucial takeaway from this is twofold. First, all States are inherently coercive insofar as they claim the right to compel the people living within their boundaries to heed and obey the commands and laws of the government, in order to realize the instrumental ends of the State, whether or not those commands and laws are rationally justified or morally right on independently ethical grounds. And second, all States are also inherently authoritarian insofar as they claim that the commands and laws issued by its government are right just because the government says that they're right and possesses the power to coerce, not because those commands or laws are rationally justified and morally right on independent ethical grounds. To the extent that the military-industrial-digital complex in fact controls all States in the contemporary 21st-century world, then it now controls coercive authoritarianism everywhere.

Against that conceptual backdrop, then what I'm calling *cybernetic civil disobedience* is essentially connected with Martin Luther King Jr's general doctrine of civil disobedience (King, 2018) as per the following nine-step argument.

1. By *violence*, I mean the use of actually or potentially destructive force, and by *nonviolence* I mean the refusal to use actually or potentially destructive force.
2. In view of broadly Kantian nonideal dignitarian moral theory, violence with respect to people is rarely if ever rationally or morally justified; indeed, except in last-resort cases of self-defense against violent attack or in order to protect the innocent from violent attack, universal nonviolence with respect to people is rationally justified and morally obligatory.
3. Nevertheless, sometimes it's not only permissible, but even rationally justified and morally obligatory, to be nonviolent with respect to people yet also violent

with respect to private property, if the relevant private property represents a basic and widespread source of violations of sufficient respect for universal human dignity—for example, if it's private property owned by technocratic global corporate capitalist conglomerates or corporations, that expresses and implements an inherently oppressive social system, such as the symbiotic combination of systemic racism, technocratic global corporate capitalism, and the coercive authoritarianism of the State (for example, of the police and the legal justice system of mass incarceration)—and the purpose of the violence with respect to private property of this kind is solely to change this inherently oppressive social system into something fundamentally better, in that it sufficiently respects universal human dignity.

4. Martin Luther King Jr (henceforth MLK), argues that massive nonviolent (with respect to people) civil disobedience is required in order to effect fundamental social change for the better in inherently oppressive social systems, and also that this nonviolent civil disobedience can include "direct action" such as the disruption of the daily operations of the inherently oppressive symbiotic social system of systemic racism, technocratic global corporate capitalism, and the coercive authoritarianism of the State, perhaps even including violence with respect to private property owned by technocratic global corporate capitalist individual magnates, conglomerates, or corporations (King, 2018).
5. Although MLK does not explicitly draw this distinction, there's nevertheless a basic difference between (5i) coercion, which is either (5ia) imposing or threatening to impose violence on people or (5ib) imposing or threatening to impose salient although nonviolent harms on people, in order to compel those people to do various things, or heed various commands or demands, in order to bring about egoistic or publicly beneficial ends of the coercer, and (5ii) noncoercion, which is the refusal to engage in coercion.
6. Since coercion treats other people as mere means or mere things, and not as persons with dignity, it violates sufficient respect for human dignity; hence all coercion is rationally unjustified and immoral, even if it's beneficial for many people.
7. So only nonviolent (with respect to people), noncoercive civil disobedience is rationally justified and morally acceptable for the purposes of effecting fundamental social change for the better in inherently oppressive social systems, and only nonviolent (with respect to people), noncoercive civil disobedient "direct action" or "disruption" is fully consistent with MLK's overall moral and political philosophy and with broadly Kantian nonideal dignitarian moral theory.
8. Therefore, nonviolent, noncoercive disobedient "direct action" or "disruption" with respect to cybernetic oppression and cybernetic totalitarianism is fully

consistent with MLK's overall moral and political philosophy, and with broadly Kantian nonideal dignitarian moral theory, alike.

9. Cybernetic civil disobedience, then, as per DCI, is the rationally and morally justified refusal to heed, or the direct violation of, rationally unjustified and immoral commands or laws of the military-industrial-digital complex about its design, creation, and use of digital technology, for the sake of universally sufficiently respecting human dignity.

In turn, I think that there are at least four modes of cybernetic civil disobedience that are rationally and morally justified by DDE and DCI: (i) cybernetic direct action or disruption, for example, strikes, marches, sit-ins, occupations, overt cyber-hacking, etc., directed against any part or phase of the military-industrial-digital complex's oppressive design, creation, and use of digital technology, for the sake of sufficiently respecting human dignity, (ii) what I call *cybernetic emancipatory free speech or freedom of expression*, that is, free speech or free expression whose essential purpose is to criticize and reject the military-industrial-digital complex's oppressive design, creation, and use of digital technology, for the sake of sufficiently respecting human dignity, (iii) cybernetic counter-cultural escape into independent, creative, meaningful activity, for example, artistic activity of all kinds, crafts of all kinds, scholarship of all kinds, especially including philosophy, and, more generally, J.S. Mill's "experiments of living" (Mill, 1978: esp. ch. III), whose essential purpose is to escape from and reject the military-industrial-digital complex's oppressive design, creation, and use of digital technology, for the sake of sufficiently respecting human dignity, and (iv) what, following the lead of the political anthropologist James C. Scott, I'll call *cybernetic weapons of the weak* (Scott, 1985), for example, foot-dragging, theft, sabotage, defacement of property, covert cyber-hacking, and above all, covert noncompliance—see section 6.10 below—directed against any part or phase of the military-industrial-digital complex's oppressive design, creation, and use of digital technology, whose essential purpose is to undermine the military-industrial-digital complex's oppressive design, creation, and use of digital technology, for the sake of sufficiently respecting human dignity.

Finally, as regards (i) to (iv) and any other modes of cybernetic civil disobedience rationally and morally justified by DCI, it's crucial to emphasize, re-emphasize, and re-re-emphasize that according to the refined version of MLK's conception of civil disobedience, together with broadly Kantian nonideal dignitarian moral theory, all cybernetic civil disobedience must always and everywhere be (i) nonviolent (with respect to people), (ii) noncoercive, and (iii) non-authoritarian. Therefore, this is most emphatically not the terrorist neo-Luddism of The Unabomber.

6.10 Cybernetic Radical Enlightenment, Stage III: Developing Our Own Innate Capacities

In this section, I'll work out the basic details of how we can individually engage in the abolition, refusal, and radical transformation of dignity-violating digital technology.

Riffing non-violently on Frank Herbert's idea of a "Butlerian Jihad" (Herbert, 1965/1990), I'll put forward the thought that in the face of cybernetic oppression, what humankind can and should do is non-violently (with respect to people) push back against our excessive reliance on and indeed addiction to digital technology, by systematically developing our own essentially embodied innate capacities through self-disciplined practice, completely independently of digital technology, to the extent that this is humanly possible. Now, using the list of seven innate capacities that I laid out in section 2.7, I'll formulate a corresponding list of fifteen modes of disciplined practice that anyone can do at home, or near home, completely independently of digital technology, to the extent that this is humanly possible. We'll thereby have a list of fifteen non-compliant cybernetic weapons of the weak (Scott, 1985) that we can individually use to push back against cybernetic oppression. To be sure, at least some of these fifteen things could be behaviorally simulated by digital technology, in the form of sophisticated robots, for example, by ChatGPT. But it's also a conceptual truth that none of those behavioral simulations, precisely because they're essentially nothing but implemented algorithms, will be done either independently of digital technology, or because they're independent of digital technology, or in order to push back against cybernetic oppression. Moreover, because I think that these fifteen modes of disciplined practice are precisely the sorts of things that radically enlightened dignitarians not only can but also should be doing, I'll frame them as categorical imperatives.

First, enjoy aesthetic experiences of all kinds, in all sense modalities. More specifically, sense-perceive the natural universe around you as if you had never perceived it before, as a non-denumerably infinite field of objects, each of which is an essentially rich actual or potential target for the subjective experiences of beauty and sublimity (Hanna, 2024b: ch. 15).

Second, explore hypnagogic states, hovering between waking and sleeping.

Third, meditate every day, even if only for a few minutes.

Fourth, fall or re-fall in love with someone or something—philosophy, for example—no matter how old you are, and more generally, actively love other people and

wholeheartedly love pursuing certain projects for their own sake. Why? Because, as Barbara Ehrenreich puts it in *Fear of Falling:*

> Love is one of the only ways we have, in a secular, commercial [and digital-technological—RH] culture, of talking about transcendent experience. (Ehrenreich, 1989)

Fifth, engage in visual sense perception and everyday bodily performances by using mirrors (see section 2.4 above).

Sixth, free-associate ideas about any topic of great interest to you, and see what creatively emerges.

Seventh, read hard-copy books, and write out your ideas in longhand. This is essential. The act or process of reading, and the act or process of longhand writing, are in fact essentially embodied, due to the egocentric centering of the living animal body of the rational human minded animal reader in 3-D orientable space for the purposes of cognizing a physical text and scanning it from left to right, or right to left, and from top to bottom, or bottom to top, and for longhand writing in order to create a physical text (see section 2.5 above).

From this, it follows that it should be empirically confirmable not only (i) that contemporary children who grow up reading and writing using physical texts and longhand writing technology (for example, pen and paper), in addition to digital texts and digital writing technology, are better readers, writers, and thinkers than those who grow up using digital texts and digital writing technology exclusively, but also (ii) that people of any age who read and write using physical texts and physical writing technology as well as digital texts and digital writing technology, are better readers, writers, and thinkers than those who use digital texts and digital writing technology exclusively. Indeed, as regards (i), this is empirically confirmed by the work of the cognitive neuroscientist Audrey van der Meer and her associates (van der Meer, Askvik, and van der Weel, 2020).

Eighth, find subject areas that interest you and then memorize things about them.

Ninth, do simple or complicated arithmetic calculations, or simple or complicated logic puzzles, in your head or on paper, or do crossword puzzles, acrostics, etc., on paper.

Tenth, study some special subject that's always fascinated you, but always seemed to be too complicated or difficult, and see what sense you can make out of it on your own.

Eleventh, go for daily walks without your smartphone or any other digital technology, then simply look at, listen to, smell, and/or touch your surroundings, and take note of changes in your local environment day-by-day.

Twelfth, spontaneously dance freestyle, just for the hell of it, no matter how silly it looks; this is also known as "the freedom dance" (Hanna, 2024b: section 2.2; see also Ehrenreich, 2006).

Thirteenth, write an autobiography, no matter how short, and then ask yourself whether your life so far has been worth living or not.

Fourteenth, if you've answered "yes" to that question, then self-consciously resolve to continue to do the things you've been doing that give your life meaning, but now by spontaneously doing them even better and more creatively than before.

Fifteenth, if you've answered "no" to that question, then self-consciously resolve to change your life for the better, and then spontaneously give it a serious try.

7 Conclusion: Dignitarian Neo-Luddism with Respect to Digital Technology

I've noted several times that the upshot of this book, as a radically robust and broadly Kantian version of ethical anthropocentrism, and as a direct expression and logical implication of the doctrine yielded by DDE and DCI, is dignitarian neo-Luddism with respect to digital technology. But precisely what is this doctrine? For clarity's sake, I'll start with some definitions. According to *The Oxford Encyclopedic Dictionary*, the term "Luddite" means this:

> 1. *hist.* a member of the bands of English craftsmen who, when their jobs were threatened by the progressive introduction of machinery into their trades in the early 19th c. attempted to reverse the trend towards mechanization by wrecking the offending machines.... 2. a person opposed to increased industrialization or new technology.... [perh. f. Ned *Lud[d]*, an insane person said to have destroyed two stocking-frames *c.* 1799] (Hawkins and Allen, 1991: p. 856)

Generalizing from that, and also precisifying a little, I'll say that either classical or modern *Luddism* says that

> all mechanical technology is bad and wrong, because it harms and oppresses ordinary people (i. e., people other than technocrats), and therefore all mechanical technology should be rejected and destroyed.

By an important contrast, however, as I'm understanding it, modern *neo-Luddism* (see, for example, Glendinning, 1990) says that

> not all mechanical technology is bad and wrong, but instead all and only the mechanical technology that harms and oppresses ordinary people (i. e., people other than technocrats) is bad and wrong, and therefore all and only this bad and wrong mechanical technology should be rejected but not—except in extreme cases of mechanical technology whose coercive use is actually violently harming and oppressing ordinary people, for example, weapons being used for mass destruction or mass murder—destroyed, rather only either simply refused, non-violently dismantled, or radically transformed into its moral opposite.

Now, as I noted in the Introduction, by *digital technology* I mean all mechanical technology that inherently involves computers, algorithms, digital data or information, artificial intelligence/AI, or robotics. Then, what I call *neo-Luddism with respect to digital technology* says that

https://doi.org/10.1515/9783111260525-010

> not all digital technology is bad and wrong, but instead all and only the digital technology that harms and oppresses ordinary people (i. e., people other than digital technocrats) is bad and wrong, and therefore all and only this bad and wrong digital technology should be rejected but not—except in extreme cases of digital technology whose coercive use is actually violently harming and oppressing ordinary people, for example, digitally-driven weapons or weapons-systems being used for mass destruction or mass murder—destroyed, rather only either simply refused, non-violently dismantled, or radically transformed into its moral opposite.

Finally, what I call *dignitarian neo-Luddism with respect to digital technology* says that

> not all digital technology is bad and wrong,[35] but instead all and only the digital technology that harms and oppresses ordinary people (i. e., people other than digital technocrats), by either failing to respect our human dignity sufficiently or by outright violating our human dignity, is bad and wrong, and therefore all and only this bad and wrong digital technology should be rejected but not—except in extreme cases of digital technology whose coercive use is actually violently harming and oppressing ordinary people, for example, digitally-driven weapons or weapons-systems being used for mass destruction or mass murder—destroyed, rather only either simply refused, non-violently dismantled, or radically transformed into its moral opposite.

So much for the definitions, and now for some moral imperatives. What I strongly believe is that we all ought to be dignitarian neo-Luddites with respect to digital technology. Why? To be sure, as we've seen in chapters 5 and 6, there are many ways in which digital technology can be bad and wrong in the dignitarian sense, including invasive digital surveillance, digitally-driven weapons and weapon systems, algorithmic bias, and digital manipulation and nudging—not to mention its making substantially contributions to disastrous climate change (Parshley, 2024). But as I anticipated this point in section 2.9 above, the principal reason for being a dignitarian neo-Luddite with respect to digital technology is that our excessive use of and indeed addiction to digital technology is systematically under-

35 It needs to be emphasized and re-emphasized that dignitarian neo-Luddism with respect to digital technology is *also* committed to the *positive* dignitarian moral doctrine that *some* digital technology is good and right, and therefore *ought to be used*, precisely because it promotes the betterment of humankind and sufficiently respects human dignity. That's directly entailed by DCI. But in this final chapter, I'm focusing on the negative dignitarian moral doctrine.

mining our innate capacities for thinking, caring, and acting for ourselves. This is preeminently true with respect to the new chatbots—LaMDA, Bard, Sydney, and ChatGPT, and others-to-be—as well as AI more generally, and also Organoid Intelligence, aka OI, at least prospectively (see section 2.10 above). But it's also to an increasingly important degree true for our excessive use of and addiction to smartphones, desktop and laptop computers, the internet, social media, and so on and so forth. Then, when you combine our excessive use of and addiction to chatbots and AI (or in the future, also OI) with our excessive use of and addiction to smartphones, desktop and laptop computers, the internet, social media, etc., the result is nothing less than an all-out existential attack on our rational human mindedness or intelligence, where by "our rational human mindedness or intelligence" I mean the essentially embodied, unified set of seven basic innate cognitive, affective, and practical capacities present in all and only those human animals possessing the essentially embodied neurobiological basis of those capacities (see section 2.7). This essentially embodied, unified set of capacities constitutes our human real personhood, which in turn is the metaphysical ground of human dignity (see chapter 3). Hence this all-out existential attack on our rational human mindedness or intelligence is also an all-out existential attack on our human dignity. In short, as I put it in section 2.9.1 above, it's *the invasion of the mind snatchers*.

Now, what is to be done? Perhaps dignitarian neo-Luddism with respect to digital technology will become a worldwide, world-changing social and political movement, comparable to the Ban-the-Bomb and anti-nuclear movement: I wholeheartedly hope so. If we ban all further giant AI experiments and LLM/chatbot technology now, when it's already obvious what their existential threat to humankind is, then the world will be a substantially better place, just as the world would have been a substantially better place if we had banned the A-bomb and nuclear weapons technology after its initial tests, when it was already obvious what their existential threat to humankind was.

But, in the meantime, given the immense power of the military-industrial-digital complex, individual dignitarian digital neo-Luddites like you and I (assuming that you agree with what I've argued in this book, that is) cannot do very much to change the world. Nevertheless, and as I proposed in section 6.10 above, updating J.C. Scott's "weapons of the weak" (Scott, 1985) for our 21st-century context, a world pervaded by digital technology and dominated by the military-industrial-digital complex, I do think that we individually can become what I'll call *Daily Dignitarian Digital Refusards*, aka DDDRs.[36] How? We can become DDDRs by resolving to cultivate our own innate capacities in a self-disciplined and autonomous way, for six

36 Pronounced "didd-ers."

waking hours every day, altogether independently of digital technology, to the extent that this is humanly possible.

To take just a few of many examples of Daily Dignitarian Digital Refusardism, you can log off all your digital devices for six waking hours a day, you can refuse to use ChatGPT or any other chatbot whatsoever, and you can instead do some or all of the fifteen things I listed in section 6.10. Even the military-industrial-digital complex, with all its immense power, can't stop you from practicing Daily Dignitarian Digital Refusardism in its many modes, each of which involves the self-disciplined, autonomous cultivation of your innate capacities. And if the day does ever come when the military-industrial-digital complex *can* stop you from being a DDDR, whether by means of permanently-implanted brain-computer interfaces or some other malign instruments of digital-technological harm and oppression, then that really and truly will be the real-world arrival of the globalized version of the totalitarian State of Alphaville, and also the end of the road for rational humankind and human dignity (for a similar line of thinking, see also Corbyn, 2023; Farhany, 2023).

And that brings me to the hortatory conclusion of this book. Since no mechanical system can ever be alive, therefore no form of digital technology, and no robot, no matter how sophisticated, can ever be alive. Therefore no form of digital technology, and no robot, no matter how sophisticated, can ever have a life of its own. Therefore no form of digital technology, and no robot, no matter how sophisticated, can ever freely change its own life for the better. But *we* can freely change our lives for the better, by engaging in the abolition, refusal, or radical transformation of human-dignity-violating digital technology.

References

Adams, D. (2002) *The Hitchhiker's Guide to the Galaxy.* New York: Ballantine/Del Rey.

Adriaans, P. (2020) "Information." In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Fall Edition. Available online at URL = <https://plato.stanford.edu/archives/fall2020/entries/information/>, last accessed 8 January 2025.

Agar, N. (2020) "How to Treat Machines That Might Have Minds." *Philosophy & Technology* 33, 2: 269 – 282.

Ahuja, A. (2000) "An Organized Death." *London Times.* 4 September.

Allen, C. and Bekoff, M. (1997) *Species of Mind.* Cambridge, MA: MIT Press.

Allen, C., Wallach, W., and Smit, I. (2011) "Why Machine Ethics?" In Anderson and Anderson, 2011: pp. 51 – 61.

Anderson, C. (2017) *White Rage: The Unspoken Truth of Our Racial Divide.* 2nd edn. London/New York: Bloomsbury.

Anderson, M. and Anderson, S.L. (2011) (eds.) *Machine Ethics.* Cambridge: Cambridge Univ. Press.

Babbage, C. (1864) *Passages from the Life of a Philosopher.* London: Longman and Co. Available online at URL = <https://www.gutenberg.org/files/57532/57532-h/57532-h.htm>, last accessed 8 January 2025.

Baiasu, S. (2011) *Kant and Sartre: Re-Discovering Critical Ethics.* London: Palgrave Macmillan.

Bawden, D. (2014) "Floridi's Information Ethics." *The Occasional Informist.* 29 July. Available online at URL = <https://theoccasionalinformationist.com/2014/07/29/floridis-information-ethics/>, last accessed 8 January 2025.

BBC News (2002) "Man Held for German 'Cannibal Killing'" (12 December 2002). Available online at URL = <http://news.bbc.co.uk/2/hi/europe/2569095.stm>, last accessed 8 January 2025.

Bearzi, M. and Stanford, C. (2008) *Beautiful Minds: The Parallel Lives of Great Apes and Dolphins.* Cambridge, MA: Harvard Univ. Press.

Benacerraf, P. (1973) "Mathematical Truth." *Journal of Philosophy* 70: 661 – 680.

Benjamin, W. (1935/1969) "The Work of Art in the Age of Mechanical Reproduction." Trans. H. Zohn. In H. Arendt (ed.), *Illuminations.* New York: Schocken, pp. 1 – 26. Available online at URL = <http://web.mit.edu/allanmc/www/benjamin.pdf>, last accessed 8 January 2025.

Block, N. (1980) (ed.) *Readings in the Philosophy of Psychology.* 2 vols. Cambridge, MA: Harvard Univ. Press, vol. 1.

Boethius (2022) "A Philosopher's Diary, #9–How Much Does the Chatbot Brouhaha Affect Anarcho-Philosophical Teaching and Learning?" *Against Professional Philosophy.* 25 December. Available online at URL = <https://againstprofphil.org/2022/12/25/a-philosophers-diary-9-how-much-does-the-chatbot-brouhaha-affect-anarcho-philosophical-teaching-and-learning/>, last accessed 8 January 2025.

Boolos, G. and Jeffrey, R. (1989) *Computability and Logic.* 3rd edn. Cambridge: Cambridge Univ. Press.

Boonin, D. (2003) *A Defense of Abortion.* Cambridge: Cambridge Univ. Press.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford Univ. Press.

Bostrom, N. and Yudkowsky, E. (2014) "The Ethics of Artificial Intelligence." In K. Frankish and W. Ramsey (eds.), *The Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge Univ. Press, pp. 316 – 334.

Bregman, R. (2020) *Humankind: A Hopeful History.* Trans. E. Manton and E. Moore. New York: Little, Brown and Co.

https://doi.org/10.1515/9783111260525-011

Brian, D. (1996) "*Einstein: A Life*, Ch 1." *The Washington Post.* Available online at URL = <https://www.washingtonpost.com/wp-srv/style/longterm/books/chap1/einstein.htm>, last accessed 8 January 2025.

Bringsjord, S., Taylor, J., van Heuveln, B., and Arkoudas, K. (2011) "Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions Are Guaranteed to Be Ethically Correct." In Anderson and Anderson, 2011: pp. 361 – 374.

Britannica (2025) "Transhumanism." Available online at URL = <https://www.britannica.com/topic/transhumanism>, last accessed 8 January 2025.

Bromell, N. (2021) *The Powers of Dignity: The Black Political Philosophy of Frederick Douglass.* Durham, NC: Duke Univ. Press.

Bryson, J. (2010) "Robots Should Be Slaves." In Y. Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues.* Natural Language Processing 8. Amsterdam: John Benjamins, pp. 63 – 74.

Bryson, J. (2019) "The Past Decade and Future of AI's Impact on Society." In *Towards a New Enlightenment: A Transcendent Decade.* Madrid: Turner-BVVA. Available online at URL = < https://static1.squarespace.com/static/5e13e4b93175437bccfc4545/t/67057b00d95cad198c5575da/1728412417139/the-past-decade-and-future-of-ai-impact-on-society-bbva.pdf >, last accessed 13 February 2025.

Burns, C., Izmailov, P., Kirchner, J.H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Sutskever, J.L.I., and Wu, J. (2023) "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." *OpenAI.* 14 December. Available online at URL = <https://cdn.openai.com/papers/weak-to-strong-generalization.pdf>, last accessed 8 January 2025.

Cahn, S. and Markie, P. (2006) (eds.) *Ethics: History, Theory, and Contemporary Issues.* 3rd edn. New York: Oxford Univ. Press, pp. 396 – 402.

Calvarly, D.J. (2011) "Legal Rights for Machines: Some Fundamental Concepts." In Anderson and Anderson, 2011: pp. 213 – 228.

Camus, A. (1956) *The Rebel: An Essay on Man in Revolt.* Trans. A. Bower. New York: Vintage.

Cantor, G. (1891) "Ueber eine elementare Frage der Mannigfaltigkeitslehre." *Jahresbericht der Deutschen Mathematiker-Vereinigung* 1: 75 – 78.

Cantor, G. (2019) "A Translation of G. Cantor's 'Ueber eine elementare Frage der Mannigfaltigkeitslehre'." Trans. P.P. Jones et al. 23 August. Available online at URL = <https://www.researchgate.net/publication/335364685_A_Translation_of_G_Cantor%27s_Ueber_eine_elementare_Frage_der_Mannigfaltigkeitslehre>, last accessed 8 January 2025.

Carroll, L. (1988) *Through the Looking-Glass.* New York: Dial.

Chalmers, D. (1996) *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford Univ. Press.

CHM (2024) "The Babbage Engine." Computer History Museum. Available online at <URL = https://www.computerhistory.org/babbage/>, last accessed 8 January 2025.

Chomsky, N. (1957) *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1988) *Language and Problems of Knowledge.* Cambridge, MA: MIT Press.

Chomsky, N., Roberts, I., and Watumull, J. (2023) "Noam Chomsky: The False Promise of ChatGPT." *New York Times.* 8 March. Available online at URL = <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>, last accessed 8 January 2025.

Clarke, R. (2011) "Asimov's Laws of Robotics: Implications for Information Technology." In Anderson and Anderson, 2011: pp. 254 – 284.

Coeckelbergh, M. (2020) *AI Ethics.* Cambridge, MA: MIT Press.

Corbyn, Z. (2023) "Prof Nita Farahany: 'We Need a New Human Right to Cognitive Liberty'." *The Guardian.* 4 March. Available online at URL = <https://www.theguardian.com/science/2023/mar/04/prof-nita-farahany-we-need-a-new-human-right-to-cognitive-liberty>, last accessed 8 January 2025.

Crary, J. (2013) *24/7: Late Capitalism and the Ends of Sleep.* London: Verso.

Dafoe, A. (2015) "On Technological Determinism: A Typology, Scope Conditions, and a Mechanism." *Science, Technology, & Human Values* 40: 1047–1076. Also available online: DOI 10.1177/0162243915579283.

Dammbeck, L. (2003) (dir) *The Net: The Unabomber, LSD, and the Internet* (*Das Netz*). Available online at URL = <https://archive.org/details/DasNetz_LutzDammbeck>, last accessed 8 January 2025.

Danto, A. (1964) "The Artworld." *Journal of Philosophy* 61: 571–584. Available online at URL = <https://www.jstor.org/stable/2022937>, last accessed 8 January 2025.

Danto, A. (1981) *The Transfiguration of the Commonplace: A Philosophy of Art.* Cambridge, MA: Harvard Univ. Press.

Data ID (2022) "What's OCR?" *Data Identification Online.* Available online at URL = <http://www.dataid.com/aboutocr.htm>, last accessed 8 January 2025.

Davis, A. (2003) *Are Prisons Obsolete?* New York: Seven Stories Press.

Dawkins, R. (2006) *The Selfish Gene.* Oxford: Oxford Univ. Press.

Depeche Mode (1990) "Policy of Truth." *Violator.* Available online at URL = <https://www.youtube.com/watch?v=M2VBmHOYpV8>, last accessed 8 January 2025.

Deresiewicz, W. (2023) "Why AI Will Never Rival Human Creativity." *Persuasion.* 8 May. Available online at URL = <https://www.persuasion.community/p/why-ai-will-never-rival-human-creativity>, last accessed 8 January 2025.

Dick, P.K. (1968) *Do Androids Dream of Electric Sheep?* New York: Doubleday.

Dietrich, E. (2011) "*Homo Sapiens* 2.0: Building the Better Robots of Our Nature." In Anderson and Anderson, 2011: pp. 531–538.

Donagan, A. (1984) "Consistency in Rationalist Moral Systems." *Journal of Philosophy* 81: 291–309.

Donagan, A. (1993) "Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy." *Ethics* 104: 7–21.

Dreyfus, H. (1972) *What Computers Can't Do.* Cambridge, MA: MIT Press.

Dreyfus, H. (1979) *What Computers Can't Do*. 2nd edn. Cambridge, MA: MIT Press.

Dreyfus, H. (1992) *What Computers Still Can't Do.* Cambridge, MA: MIT Press.

Dreyfus, H. and Dreyfus, S. (1986) *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* Oxford: Blackwell.

Dyson, D. (2012) *Turing's Cathedral.* New York: Vintage.

Dyson, F. (2008) *The Scientist as Rebel.* New York: NRYB Collections.

Ehrenreich, B. (1989) *Fear of Falling: The Inner Life of the Middle Class.* New York: Pantheon. Available online at URL = <https://archive.org/details/fearoffalling00barb/page/n5/mode/2up>, last accessed 8 January 2025.

Ehrenreich, B. (2006) *Dancing in the Streets: A History of Collective Joy.* New York: Metropolitan Books. Available online at URL = <https://archive.org/details/dancinginstreets00ehre>, last accessed 8 January 2025.

Ekirch, A.R. (2001) "Sleep We Have Lost: Pre-Industrial Slumber in the British Isles." *American Historical Review* 106: 343–386.

Ekirch, A.R. (2005) *At Day's Close: Night in Times Past.* New York: W.W. Norton.

Ekirch, A.R. (2016) "Segmented Sleep in Preindustrial Societies." *SLEEP* 39: 715–716.

Else, J. (1980) (dir.) *The Day After Trinity: J. Robert Oppenheimer and the Atomic Bomb.* Pyramid Films, PBS (television).

Emerson, R.W. (1957) "Self-Reliance." In S. E. Whicher (ed.), *Selections from Ralph Waldo Emerson.* Boston, MA: Houghton Mifflin, pp. 147 – 168.

Equal Justice Initiative (2024) "Bryan Stevenson." Available online at URL = <https://eji.org/bryan-stevenson/>, last accessed 8 January 2025.

Etinson, A. (2020) "What's So Special about Human Dignity?" *Philosophy and Public Affairs* 48: 353 – 381.

Farhany, N. (2023) *The Battle for the Brain.* New York: St Martin's Press.

Fisher, M. (2013) "Exiting the Vampire Castle." *Open Democracy UK*. 24 November. Available online at URL = <https://www.opendemocracy.net/ourkingdom/mark-fisher/exiting-vampire-castle>, last accessed 8 January 2025.

Floridi, L. (2011) "On the Morality of Artificial Agents." In Anderson and Anderson, 2011: pp. 184 – 212.

Floridi, L. (2013) *The Ethics of Information.* Oxford: Oxford Univ. Press.

Floridi, L. (2016) "Should We Be Afraid of AI? Machines Seem to Be Getting Smarter and Smarter and Much Better at Human Jobs, Yet True AI Is Utterly Implausible. Why?" *Aeon.* 9 May. Available online at URL = <https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible>, last accessed 8 January 2025.

Fong, B.Y. (2018) "Log Off." *Jacobin.* 29 November. Available online at URL = <https://jacobinmag.com/2018/11/log-off-facebook-twitter-social-media-addiction>, last accessed 8 January 2025.

Frankfurt, H. (1988a) *The Importance of What We Care About.* Cambridge: Cambridge Univ. Press.

Frankfurt, H. (1988b) "Freedom of the Will and the Concept of a Person." In Frankfurt, 1988a: pp. 11 – 25.

Frankfurt, H. (1988c) "Identification and Externality." In Frankfurt, 1988a: pp. 58 – 68.

Frankfurt, H. (1988d) "The Importance of What We Care About." In Frankfurt, 1988a: pp. 80 – 94.

Frankfurt, H. (1988e) "Identification and Wholeheartedness." In Frankfurt, 1988a: 159 – 176.

Frankfurt, H. (2015) *On Inequality.* Princeton, NJ: Princeton Univ. Press.

French, P. (1979) "The Corporation as a Moral Person." *American Philosophical Quarterly* 16: 207 – 215.

Fromm, E. (1961) *Marx's Concept of Man.* New York: Frederick Ungar.

Future of Life Institute (2023) "Pause Giant AI Experiments: An Open Letter." 22 March. Available online at URL = <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>, last accessed 8 January 2025.

Gare, A. (2016) "Beyond Modernism and Postmodernism: The Narrative of Re-Embodiments." In R. Thomas-Pellicer, V. De Lucia, and S. Sullivan (eds.), *Contributions to Law, Philosophy and Ecology: Exploring Re-Embodiments.* London: Routledge, ch. 1, pp.27 – 46. Also available online at URL = <https://www.academia.edu/67052403/Beyond_Modernism_and_Postmodernism_The_Narrative_of_Re_embodiments_>, last accessed 8 January 2025.

Gips, J. (2011) "Towards the Ethical Robot." In Anderson and Anderson, 2011: pp. 244 – 253.

Glendinning, C. (1990) "Notes Toward a Neo-Luddite Manifesto." *The Anarchist Library.* Available online at URL = <https://theanarchistlibrary.org/library/chellis-glendinning-notes-toward-a-neo-luddite-manifesto>, last accessed 8 January 2025.

Godard, J.-L. (dir). (1965) *Alphaville.* Paris: Chaumiane Productions, Filmstudio.

Gödel, K. (1931/1967) "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems." In van Heijenoort, 1967: pp. 596 – 617.

Goldstine, H. (1980) *The Computer from Pascal to von Neumann.* Princeton, NJ: Princeton Univ. Press.

Gordon, J.-S. and Nyholm, S. (2021) “Ethics of Artificial Intelligence.” *Internet Encyclopedia of Philosophy.* Available online at URL = <https://iep.utm.edu/ethics-of-artificial-intelligence/>, last accessed 8 January 2025.

Graeber, D. (2013) “On the Phenomenon of Bullshit Jobs: A Work Rant.” *Strike* 3. Available online at URL = <http://strikemag.org/bullshit-jobs/>, last accessed 8 January 2025.

Graeber, D. (2018) *Bullshit Jobs: A Theory.* New York: Simon & Schuster.

Grau, C. (2011) “There Is No ‘I’ in ‘Robot’: Robots and Utilitarianism.” In Anderson and Anderson, 2011: pp. 451 – 463.

Griffin, D.R. (1976) *The Question of Animal Awareness.* New York: Rockefeller Univ. Press.

Griffin, D.R. (1984) *Animal Thinking.* Cambridge, MA: Harvard Univ. Press.

Griffin, D.R. (2001) *Animal Minds.* Chicago, IL: University of Chicago Press.

Griffiths, M.D (2018) “Addicted to Social Media?” *Psychology Today.* 7 May. Available online at URL = <https://www.psychologytoday.com/us/blog/in-excess/201805/addicted-social-media>, last accessed 8 January 2025.

Guarini, M. (2011) “Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism–Generalism Debate.” In Anderson and Anderson, 2011: pp. 316 – 334.

Gunkel, D.J. (2012) *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* Cambridge, MA: MIT Press.

Hall, J.S. (2011) “Ethics for Machines.” In Anderson and Anderson, 2011: pp. 28 – 44.

Hanna, R. (2001) *Kant and the Foundations of Analytic Philosophy.* Oxford: Clarendon/Oxford Univ. Press.

Hanna, R. (2006a) *Kant, Science, and Human Nature.* Oxford: Clarendon/Oxford Univ. Press.

Hanna, R. (2006b) *Rationality and Logic.* Cambridge, MA: MIT Press.

Hanna, R. (2006c) “Rationality and the Ethics of Logic.” *Journal of Philosophy* 103: 67 – 100.

Hanna, R. (2009) “Freedom, Teleology, and Rational Causation.” *Kant Yearbook* 1: 99 – 142.

Hanna, R. (2011a) “Minding the Body.” *Philosophical Topics* 39: 15 – 40.

Hanna, R. (2011b) “What Is the Self?” *Annals of the New York Academy of Sciences* 1234: 121 – 123.

Hanna, R. (2015a) *Cognition, Content, and the A Priori: A Study in the Philosophy of Mind and Knowledge.* The Rational Human Condition, Vol. 5. Oxford: Oxford Univ. Press.

Hanna, R. (2015b) “Transcendental Normativity and the Avatars of Psychologism.” In A. Stati (ed.), *Husserl’s Ideas I: New Commentaries and Interpretations.* Berlin: De Gruyter, pp. 51 – 67.

Hanna, R. (2015c) “A World with Persons but Without Borders.” *Oxford University Press Blog.* 11 October. Available online at URL = <https://blog.oup.com/2015/10/kant-global-refugee-crisis/>, last accessed 8 January 2025.

Hanna, R. (2016) “Radical Enlightenment: Existential Kantian Cosmopolitan Anarchism, with a Concluding Quasi-Federalist Postscript.” In D. Heidemann and K. Stoppenbrink (eds.), *Join, Or Die: Philosophical Foundations of Federalism.* Berlin: De Gruyter, pp. 63 – 90.

Hanna, R. (2017a) “Kant’s Theory of Judgment.” In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Winter Edition. Available online at URL = <https://plato.stanford.edu/archives/win2017/entries/kant-judgment/>, last accessed 8 January 2025.

Hanna, R. (2017b) “Kant, Adorno, and Autonomy.” *Critique* (blog). Also available online in preview at URL = <https://www.academia.edu/44881739/Kant_Adorno_and_Autonomy_Critique_2017_>, last accessed 8 January 2025.

Hanna, R. (2017c) “Exiting the State and Debunking the State of Nature.” *Con-Textos Kantianos* 5: 167 – 189. Available online at URL = < https://revistas.ucm.es/index.php/KANT/article/view/89975/4564456565918>, last accessed 13 February 2025.

Hanna, R. (2018a) *Preface and General Introduction, Supplementary Essays, and General Bibliography.* The Rational Human Condition, Vol. 1. New York: Nova Science.

Hanna, R. (2018b) *Deep Freedom and Real Persons: A Study in Metaphysics.* The Rational Human Condition, Vol. 2. New York: Nova Science.

Hanna, R. (2018c) *Kantian Ethics and Human Existence: A Study in Moral Philosophy.* The Rational Human Condition, Vol. 3. New York: Nova Science.

Hanna, R. (2018d) *Kant, Agnosticism, and Anarchism: A Theological-Political Treatise.* The Rational Human Condition, Vol. 4. New York: Nova Science.

Hanna, R. (2018e) "*Kant and Cosmopolitanism* Reconsidered." *Critique* (blog). Available online in preview at URL = <https://www.academia.edu/44881351/_Kant_and_Cosmopolitanism_Reconsidered_Critique_2018_>, last accessed 8 January 2025.

Hanna, R. (2020) "The Kant Wars and the Three Faces of Kant." *Contemporary Studies in Kantian Philosophy* 5: 73 – 94. Available online at URL = <https://www.cckp.space/single-post/2020/06/15/CSKP5-2020-The-Kant-Wars-and-The-Three-Faces-of-Kant>, last accessed 8 January 2025.

Hanna, R. (2020b) "Will-Power: Essentially Embodied Agentive Phenomenology, by Way of O'Shaughnessy." In C. Erhard and T. Keiling (eds.), *Routledge Handbook: The Phenomenology of Agency.* London: Routledge, pp. 312 – 333.

Hanna, R. (2021a) *The Fate of Analysis: Analytic Philosophy from Frege to the Ash-Heap of History.* New York: Mad Duck Coalition.

Hanna, R. (2021b) "Our Sociable Sociality: A Postscript to *The Mind-Body Politic.*" *Borderless Philosophy* 4: 57 – 96. Available online at URL = <https://www.cckp.space/single-post/bp4-2021-robert-hanna-our-sociable-sociality-a-postscript-to-the-mind-body-politic>, last accessed 8 January 2025.

Hanna, R. (2023a) "Essentially Embodied Kantian Selves and The Fantasy of Transhuman Selves." *Studies in Transcendental Philosophy* 3. Available online at URL = <https://ras.jes.su/transcendental/s271326680021060-6-1-en>, last accessed 8 January 2025.

Hanna, R. (2023b) "Dignitarian Post-Capitalism." *Borderless Philosophy* 6: 99-129. Available online at URL = <https://www.cckp.space/single-post/bp6-2023-robert-hanna-dignitarian-post-capitalism-99-129>, last accessed 8 January 2025.

Hanna, R. (2024a) "Gun Crazy: A Moral Argument for Gun Abolitionism." Unpublished MS. Available online at URL = <https://www.academia.edu/61516955/Gun_Crazy_A_Moral_Argument_For_Gun_Abolitionism_March_2024_version_>, last accessed 8 January 2025.

Hanna, R. (2024b) *Science for Humans: Mind, Life, the Formal-&-Natural Sciences, and a New Concept of Nature.* New York: Springer Nature.

Hanna, R. and Maiese, M. (2009) *Embodied Minds in Action.* Oxford: Oxford Univ. Press.

Hanna, R. and Paans, O. (2020) "This Is the Way the World Ends: A Philosophy of Civilization Since 1900, and A Philosophy of the Future." *Cosmos & History* 16, 2: 1 – 53. Available online at URL = <https://cosmosandhistory.org/index.php/journal/article/view/865>, last accessed 8 January 2025.

Hanna, R. and Paans, O. (2021) "Thought-Shapers." *Cosmos & History* 17, 1: 1 – 72. Available online at URL = <http://cosmosandhistory.org/index.php/journal/article/view/923>, last accessed 8 January 2025.

Hanna, R. and Paans, O. (2022) "Creative Piety and Neo-Utopianism: Cultivating Our Global Garden." *Cosmos and History* 18, 1: 1 – 82. Available online at URL = <https://cosmosandhistory.org/index.php/journal/article/view/1017>, last accessed 8 January 2025.

Harris, S. (2012) *Free Will.* New York: Free Press.

Hawkins, J.M. and Allen, R. (1991) (eds.) *The Oxford Encyclopedic English Dictionary.* Oxford: Clarendon/Oxford Univ. Press.

Herbert, F. (1965/1990) *Dune.* New York: Ace.

Hill, T. (2002) "Moral Dilemmas, Gaps, and Residues." In T. Hill, *Human Welfare and Moral Worth: Kantian Perspectives.* Oxford: Clarendon/Oxford Univ. Press, pp. 362 – 402.

The Hammond Times (1957) "Work with New Electronic 'Brains' Opens Field for Army Math Experts." Available online at URL = <https://www.newspapers.com/clip/50687334/the-times/>, last accessed 8 January 2025.

Horn, M.B. and Jackson, C.J. (2021) "A Paradox No More: Investing in Automation and People." *MITSloan Management Review.* 22 March. Available online at URL = <https://sloanreview.mit.edu/article/a-paradox-no-more-investing-in-automation-and-people/>, last accessed 8 January 2025.

Hume, D. (1978) *Treatise of Human Nature.* 2nd edn. Oxford: Clarendon/Oxford Univ. Press.

Israel, J. (2001) *Radical Enlightenment: Philosophy and the Making of Modernity 1650 – 1750.* Oxford: Oxford Univ. Press.

Jacquette, D. (1991) "Moral Dilemmas, Disjunctive Obligations, and Kant's Principle That 'Ought' Implies 'Can'." *Synthese* 88: 43 – 55.

Jauhar, S. (2023) "When My Father Got Alzheimer's, I Had to Learn to Lie to Him." *New York Times.* 7 April. Available online at URL = <https://www.nytimes.com/2023/04/07/opinion/alzheimers-dementia-lying.html>, last accessed 8 January 2025.

Keller, A. (2023) "Artificial, But Not Intelligent: A Critical Analysis of AI and AGI." *Against Professional Philosophy.* 5 March. Available online at URL = <https://againstprofphil.org/2023/03/05/artificial-but-not-intelligent-a-critical-analysis-of-ai-and-agi/>, last accessed 8 January 2025.

Kim, J. (2011) *Philosophy of Mind.* 3rd edn. Boulder, CO: Westview.

King Jr, M.L. (2018) "Nonviolence and Social Change." *Jacobin.* 4 April. Available online at URL = <https://www.jacobinmag.com/2018/04/martin-luther-king-jr-nonviolence-direct-action>, last accessed 8 January 2025.

Kleingeld, P. (2012) *Kant and Cosmopolitanism.* Cambridge: Cambridge Univ. Press.

Korsgaard, C. (1996) "The Right to Lie: Kant on Dealing with Evil." In C. Korsgaard, *Creating the Kingdom of Ends.* Cambridge: Cambridge Univ. Press, pp. 133 – 158.

Kranzberg, M. (1981) "Work, Organization of." In W.E. Preece et al. (eds.), *Encyclopedia Britannica.* 15th edn. 30 vols. USA: Encyclopedia Britannica Inc. Macropedia, vol. 19, pp. 932 – 942.

Kripke, S. (1975) "Outline of a Theory of Truth." *Journal of Philosophy* 72: 690 – 715.

Kripke, S. (1982) *Wittgenstein on Rules and Private Language.* Cambridge, MA: Harvard Univ. Press.

Kuhse, H. and Singer, P. (1990) "Individuals, Humans, and Persons: The Issue of Moral Status." In P. Singer, H. Kuhse, S. Buckle, K. Dawson, and P. Kasimba (eds.), *Embryo Experimentation.* Cambridge: Cambridge Univ. Press, pp. 65 – 75.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Laker, B., Godley, W., Patel, C., and Cobb, D. (2020) "How to Monitor Remote Workers—Ethically." *MIT Sloan Management Review.* 2 November. Available online at URL = <https://sloanreview.mit.edu/article/how-to-monitor-remote-workers-ethically/>, last accessed 8 January 2025.

Landgrebe, J. and Smith, B. (2022) *Why Machines Will Never Rule the World: Artificial Intelligence Without Fear.* London: Routledge.

Loh, J. (2019) *Roboterethik: Eine Einführung.* Frankfurt am Main: Suhrkamp.

Lovely, G. (2024) "Can Humanity Survive AI?" *Jacobin* 52 (Winter), pp. 66 – 79. Also available online at URL = <https://jacobin.com/2024/01/can-humanity-survive-ai/>, last accessed 8 January 2025.

Mackworth, A.K. (2011) “Architectures and Ethics for Robots: Constraint Satisfaction as a Unitary Design Framework.” In Anderson and Anderson, 2011: pp. 335 – 360.

Maiese, M. (2011) *Embodiment, Emotion, and Cognition.* London: Palgrave Macmillan.

Maiese, M. and Hanna, R. (2019) *The Mind-Body Politic.* London: Palgrave Macmillan.

Maiese, M. et al. (2023) “The Shape of Lives to Come.” *Frontiers in Psychology Research Topics.* Available online at URL = <https://www.frontiersin.org/research-topics/25439/the-shape-of-lives-to-come>, last accessed 8 January 2025.

Marcus, G. and Davis, E. (2020) *Rebooting AI: Building Artificial Intelligence We Can Trust.* New York: Vintage.

Marcus, R.B. (1980) “Moral Dilemmas and Consistency.” *Journal of Philosophy* 77: 121 – 136.

Marx, K. (1961) “Economic and Philosophical Manuscripts.” Trans. T.B. Bottomore. In Fromm, 1961: pp. 90 – 196.

Marx, K. (1964) *Selected Writings in Sociology & Social Philosophy.* Trans. T.B. Bottomore. New York: McGraw-Hill.

Mason, H.E. (1996) (ed.) *Moral Dilemmas and Moral Theory.* Oxford: Oxford Univ. Press.

McDermott, D. (2011) “What Matters to a Machine?” In Anderson and Anderson, 2011: pp. 88 – 114.

McDonald, A. et al. (2021) *Developing an Ethical Technology Mindset.* MIT Sloane Management Review Executive Guides. Boston, MA: Deloitte.

McGinn, C. (1989) “Can We Solve the Mind-Body Problem?” *Mind* 98: 349 – 366.

McKeever, M. (2021) “Superintelligence and Moral Blindness.” *Medium.* 22 August. Available online at URL = <https://mittmattmutt.medium.com/superintelligence-and-moral-blindness-7436300fcb1f>, last accessed 8 January 2025.

McLaren, B.M. (2011) “Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions.” In Anderson and Anderson, 2011: pp. 297 – 315.

McWhorter, L. (2017) “The Morality of Corporate Persons.” *Southern Journal of Philosophy* 55: 126 – 148.

Metz, C. (2023) “‘The Godfather of A.I.’ Leaves Google and Warns of Danger Ahead.” *The New York Times.* 1 May. Available online at URL = <https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>, last accessed 8 January 2025.

Mill, J.S. (1978) *On Liberty.* Indianapolis IN: Hackett.

Misselhorn, C. (2019) *Grundfragen der Maschinenethik.* 4th edn. Stuttgart: Reclam.

Moor, J.H. (2011) “The Nature, Importance, and Difficulty of Machine Ethics.” In Anderson and Anderson, 2011: pp. 13 – 20.

Muldoon, J. (2021) “Facebook Is Now Meta. And It Wants to Monetize Your Whole Existence.” *Jacobin.* 28 October. Available online at URL = <https://jacobinmag.com/2021/10/mark-zuckerberg-meta-facebook-rebrand-metaverse?mc_cid=55af04bf85&mc_eid=53b430cce0>, last accessed 8 January 2025.

Müller, V.C. (2023) “Ethics of Artificial Intelligence and Robotics.” In E.N. Zalta and U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy.* Fall Edition. Available online at URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>, last accessed 8 January 2025.

Newen, A., De Bruin, L., and Gallagher, S. (2018) (eds.), *The Oxford Handbook of 4E Cognition.* Oxford: Oxford Univ. Press.

Nguyen, C.T. (2018) “Escape the Echo Chamber: Why It’s as Hard to Escape an Echo Chamber as It Is to Flee a Cult.” *Aeon.* 9 April. Available online at URL = <https://aeon.co/essays/why-its-as-hard-to-escape-an-echo-chamber-as-it-is-to-flee-a-cult>, last accessed 8 January 2025.

Nussbaum, M. (1986) *The Fragility of Goodness.* Cambridge: Cambridge Univ. Press.

Oakes, K. (2023) "Why Some People Can't Tell Left from Right." *BBC Future.* 12 January. Available online at URL = <https://www.bbc.com/future/article/20230112-why-some-people-cant-tell-left-from-right>, last accessed 8 January 2025.

O'Connell, M. (2018) "Five Books to Understand Transhumanism." *The Guardian.* 10 May. Available online at URL = <https://www.theguardian.com/books/2018/may/10/mark-oconnell-five-books-to-understand-transhumanism>, last accessed 8 January 2025.

O'Neill, O. (1989) *Constructions of Reason.* Cambridge: Cambridge Univ. Press.

O'Neill, O. (2006) "Consistency in Action." In Cahn and Markie, 2006: pp. 541 – 558.

OpenAI (2025a) OpenAI Blog. Available online at URL = <https://openai.com/blog/chatgpt/>, last accessed 8 January 2025.

OpenAI (2025b) "Research: Weak-to-Strong Generalization." Available online at URL = <https://openai.com/index/weak-to-strong-generalization/>, last accessed 8 January 2025.

Oppy, G. and Dowe, D. (2021) "The Turing Test." In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Winter Edition. Available online at URL = <https://plato.stanford.edu/archives/win2021/entries/turing-test/>, last accessed 8 January 2025.

Ostrom, E. (1990) *Governing the Commons.* Cambridge: Cambridge Univ. Press.

P2P (2025) *The Peer-to-Peer Foundation.* Available online at URL = <https://p2pfoundation.net/>, last accessed 8 January 2025.

Parshley, L. (2024) "The Hidden Environmental Impact of AI." *Jacobin.* 20 June. Available online at URL = <https://jacobin.com/2024/06/ai-data-center-energy-usage-environment?mc_cid=cc06004feb&mc_eid=53b430cce0>, last accessed 8 January 2025.

Pascal, B. (1995) *Pensées.* Trans. A.J. Krailsheimer. Rev. edn. London: Penguin.

PMA (2025) Warhol, A. "Brillo Boxes." *Philadelphia Museum of Art.* Available online at URL = <https://philamuseum.org/collection/object/89204>, last accessed 8 January 2025.

Popova, M. (2017) "The Binary Code of Body and Spirit: Computing Pioneer Alan Turing on Mortality." *The Marginalian.* 17 February. Available online at URL = <https://www.themarginalian.org/2017/02/17/alan-turing-morcom-letters/>, last accessed 8 January 2025.

Powers, T.M. (2011) "Prospects for a Kantian Machine." In Anderson and Anderson, 2011: pp. 464 – 475.

Putnam, H. (1981) *Reason, Truth, and History.* Cambridge: Cambridge Univ. Press.

Putnam, H. (1983) "There Is at Least One A Priori Truth." In H. Putnam, *Realism and Reason: Philosophical Papers, Vol. 3.* Cambridge: Cambridge Univ. Press, pp. 98 – 114.

Rachels, J. and Rachels, S. (2012) *The Elements of Moral Philosophy.* 7th edn. New York: McGraw-Hill.

Rapaport, M. (2011) (dir.) *Beats, Rhymes & Life: The Travels of a Tribe Called* Quest. Sony Pictures Classics.

Rawls, J. (1971) *A Theory of Justice.* Cambridge, MA: Harvard Univ. Press.

Ross, W.D. (1930) *The Right and the Good.* Oxford: Oxford Univ. Press.

Royce, J. (1970) *The Letters of Josiah Royce.* Chicago, IL: Univ. of Chicago Press.

Russell, B. (1940) *An Inquiry into Meaning and Truth.* London: George Allen and Unwin.

Sartre, J-P. (2006) "Existentialism Is a Humanism." Trans. B. Frechtman. In Cahn and Markie, 2006: pp. 396 – 402.

Savage-Rumbaugh, S. and Lewin, R. (1994) *Kanzi: The Ape at the Brink of the Human Mind.* New York: Wiley.

Sayre-McCord, J. (2013) "A Moral Argument Against Moral Dilemmas." Unpublished MS. Available online at URL= <https://philosophy.unc.edu/wp-content/uploads/sites/122/2013/10/A-Moral-Argument-Against-Moral-Dilemmas.pdf>, last accessed 8 January 2025.

Schapiro, T. (2006) "Kantian Rigorism and Mitigating Circumstances." *Ethics* 117: 32 – 57.
Schopenhauer, A. (1851/2004) "On the Sufferings of the World." In A. Schopenhauer, *The Essays of Arthur Schopenhauer: Studies in Pessimism.* Trans. T. Bailey Saunders. Blackmask's Library. Available online at URL = <https://spiritual-minds.com/philosophy/assorted/Philosophy%20-%20Arthur%20Schopenhauer%20-%20Studies%20In%20Pessimism.pdf>, last accessed 8 January 2025.
Schrödinger, E. (1935/1980) "The Present Situation in Quantum Mechanics." Trans. J.D. Trimmer. *Proceedings of the American Philosophical Society* 124: 323 – 338.
Schuessler, J. (2013) "Philosophy That Stirs the Waters." *New York Times.* 29 April. Available online at URL = <http://www.nytimes.com/2013/04/30/books/daniel-dennett-author-of-intuition-pumps-and-other-tools-for-thinking.html?emc=eta1&_r=0>, last accessed 8 January 2025.
Schulson, M. (2015) "User Behavior: If the Internet Is Addictive, Why Don't We Regulate It?" *Aeon.* 24 November. Available online at URL = <https://aeon.co/essays/if-the-internet-is-addictive-why-don-t-we-regulate-it>, last accessed 8 January 2025.
Scott, J.C. (1985) *Weapons of the Weak: Everyday Forms of Peasant Resistance.* New Haven, CT: Yale Univ. Press.
Scott, J.C. (2017) *Against the Grain: A Deep History of the Earliest States.* New Haven, CT: Yale Univ. Press.
Searle, J. (1980a) "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417 – 424.
Searle, J. (1980b) "Intrinsic Intentionality." *Behavioral and Brain Sciences* 3: 450 – 456.
Searle, J. (1984) *Minds, Brains, and Science.* Cambridge, MA: Harvard Univ. Press.
Searle, J. (1992) *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.
Seville, H. and Field, D.G. (2011) "What Can AI Do for Ethics?" In Anderson and Anderson, 2011: pp. 499 – 511.
Shadbolt, N. and Hampson, R. (2024) *As If Human: Ethics and Artificial Intelligence.* New Haven, CT: Yale Univ. Press.
Shafer-Landau, R. (2015) *The Fundamentals of Ethics.* 3rd edn. Oxford: Oxford Univ. Press.
Shelley, M. (1818) *Frankenstein; Or, the Modern Prometheus.* Available online at URL = <https://romantic-circles.org/editions/frankenstein/editions.2009.frankenstein>, last accessed 13 February 2025.
Shiffrin, S. (2016) *Speech Matters: On Lying, Morality, and the Law.* Princeton, NJ: Princeton Univ. Press.
Siebert, C. (2014) "Should a Chimp Be Able to Sue Its Owner?" *New York Times.* 23 April. Available online at URL = <https://www.nytimes.com/2014/04/27/magazine/the-rights-of-man-and-beast.html?emc=eta1&_r=0>, last accessed 8 January 2025.
Sipser, M. (2013) *Introduction to the Theory of Computation.* 3rd edn. Boston, MA: Cengage Learning.
Sloane, M. (2021) "The Algorithmic Auditing Trap." *OneZero.* 16 March. Available online at URL = <https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d>, last accessed 8 January 2025.
Smirnova, L. et al. (2023) "Organoid Intelligence (OI): The New Frontier in Biocomputing and Intelligence-in-a-Dish." *Frontiers in Science.* 28 February. Available online at URL = <https://www.frontiersin.org/journals/science/articles/10.3389/fsci.2023.1017235>, last accessed 8 January 2025.
Smith-Garcia, D. (2020) "Einstein Syndrome: Characteristics, Diagnosis, and Treatment." *Healthline.* Available online at URL = <https://www.healthline.com/health/einstein-syndrome>, last accessed 8 January 2025.
Sowell, T. (1997) *The Einstein Syndrome: Bright Children Who Talk Late.* New York: Basic Books.
Sullins, J.P. (2011) "When Is a Robot a Moral Agent?" In Anderson and Anderson, 2011: pp. 151 – 161.

Swedenborg Foundation (2025) "Who Was Swedenborg?" Available online at URL = <https://swedenborg.com/emanuel-swedenborg/>, last accessed 8 January 2025.

Tarski, A. (1943) "The Semantic Conception of Truth and the Foundations of Semantics." *Philosophy and Phenomenological Research* 4: 342 – 360.

Tarski, A. (1956) "The Concept of Truth in Formalized Languages." In A. Tarski, *Logic, Semantics, and Metamathematics.* Oxford: Oxford Univ. Press, pp. 152 – 278.

TM (2025) Duchamp, M. "Fountain." *Tate Modern.* Available online at URL = <https://www.tate.org.uk/art/artworks/duchamp-fountain-t07573>, last accessed 8 January 2025.

Tomasello, M. and Call, J. (1997) *Primate Cognition.* New York: Oxford Univ. Press.

Torday, J., Miller Jr, W.B., and Hanna, R. (2020) "Singularity, Life, and Mind: New Wave Organicism." In J. Torday and W.B. Miller Jr, *The Singularity of Nature.* Cambridge: Royal Society of Chemistry, ch. 20, pp. 206 – 246.

Torrance, S. (2011) "Machine Ethics and the Idea of a More-Than-Human World." In Anderson and Anderson, 2011: pp. 115 – 137.

Turilli, M. (2011) "Ethical Protocols Design." In Anderson and Anderson, 2011: pp. 375 – 397.

Turing, A. (1936/1937) "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 42: 230 – 265, with corrections in 43: 644 – 546.

Turing, A. (1950) "Computing Machinery and Intelligence." *Mind* 59: 433 – 460.

Turkle, S. (2009) *Simulation and Its Discontents.* Cambridge, MA: MIT Press.

Unamuno, M. (1954) *Tragic Sense of Life.* Trans. J.E. Crawford Flitch. New York: Dover. Available online at URL = <http://www.gutenberg.org/files/14636/14636-h/14636-h.htm>, last accessed 8 January 2025.

Vallor, S. (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting.* Oxford: Oxford Univ. Press.

van der Meer, A.L.H, Askvik, E.O., and van der Weel, F.R. (2020) "The Importance of Cursive Handwriting over Typewriting for Learning in the Classroom: A High-Density EEG Study of 12-Year-Old Children and Young Adults." *Frontiers in Psychology.* 28 July. Available online at URL = <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01810/full>, last accessed 8 January 2025.

van Heijenoort, J. (1967) (ed.) *From Frege to Gödel.* Cambridge, MA: Harvard Univ. Press.

Vitali, A. (2017) *The End of Policing.* London: Verso.

Weber, M. (1994) "The Profession and Vocation of Politics." In P. Lassman and R. Spiers (eds.), *Weber: Political Writings.* Cambridge: Cambridge Univ. Press, pp. 309 – 369.

Whitby, B. (2011) "On Computable Morality: An Examination of Machines as Moral Advisors." In Anderson and Anderson, 2011: pp. 138 – 150.

Wikipedia (2025a) "ChatGPT." Available online at URL = <https://en.wikipedia.org/wiki/ChatGPT>, last accessed 13 February 2025.

Wikipedia (2025b) "Military-Industrial Complex." Available online at URL = <https://en.wikipedia.org/wiki/Military%E2%80%93industrial_complex>, last accessed 13 February 2025.

Wikipedia (2025c) "Transhumanism." Available online at URL = <https://en.wikipedia.org/wiki/Transhumanism>, last accessed 13 February 2025.

Wikipedia (2025d) "The Turing Test: Google LaMDA." Available online at URL = <https://en.wikipedia.org/wiki/Turing_test#Google_LaMDA_chatbot>, last accessed 13 February 2025.

Wikipedia (2025e) "Rage Against the Machine." Available online at URL = <https://en.wikipedia.org/wiki/Rage_Against_the_Machine>, last accessed 13 February 2025.

Wikipedia (2025f) “Brandon Bryant (Whistleblower).” Available online at URL = <https://en.wikipedia.org/wiki/Brandon_Bryant_(whistleblower)>, last accessed 13 February 2025.
Williams, B. (1973a) “Ethical Consistency.” In B. Williams, *Problems of the Self.* Cambridge: Cambridge Univ. Press, pp. 166 – 186.
Williams, B. (1973b) “A Critique of Utilitarianism.” In J.J.C. Smart and B. Williams, *Utilitarianism: For and Against.* Cambridge: Cambridge Univ. Press, pp. 77 – 150.
Williams, B. (1981a) *Moral Luck.* Cambridge: Cambridge Univ. Press.
Williams, B. (1981b) “Moral Luck.” In Williams, 1981a: pp. 20 – 29.
Williams, B. (1981c) “Conflicts of Values.” In Williams, 1981a: pp. 71 – 82.
Wittgenstein, L. (1953) *Philosophical Investigations.* Trans. G.E.M. Anscombe. New York: Macmillan.
Yetish, G. et al. (2015) “Natural Sleep and Its Seasonal Variations in Three Pre-Industrial Societies.” *Current Biology* 25: 2862 – 2868.
Zermelo, E. (1930) “Über Grenzzahlen und Mengenbereiche.” *Fundamenta Mathematicae* 16: 29 – 47.
Zermelo, E. (1967a) “Proof That Every Set Can Be Well-Ordered.” In van Heijenoort, 1967: pp. 139 – 141.
Zermelo, E. (1967b) “A New Proof of the Possibility of a Well-Ordering.” In van Heijenoort, 1967: pp. 183 – 198.
Zermelo, E. (1967c) “Investigations in the Foundations of Set Theory I.” In van Heijenoort, 1967: pp. 199 – 215.

Index

abductive reasoning 53
absolutely universal and objective moral meta-principles 6 f., 11 f., 23
abstract objects 40
Adams, Douglas 13
addiction to digital technology 80 f., 183, 187
Adriaans, P. 1
affective framing 95
Agar, N. 155
AHC 52 – 57
Ahuja, A. 3
Algerian War 2
algorithms 1, 6, 9 – 11, 17 f., 28, 44, 57, 66, 157, 183, 186
Allen, C. 1, 33, 35, 97, 155, 186
Allen, R. 1, 33, 35, 97, 155, 186
Alpha-60 1 f.
Alphaville 1 f., 177, 189
Anderson, C. 8, 155
Anderson, M. 8, 155
Anderson, S.L. 8, 155
Anna Karenina 41
anthropocentrism 4, 11, 149 – 151, 186
Apple 77 f.
arguments 5, 17 f., 44 f., 47 – 49, 53, 56, 66, 68 f., 81, 85, 155
Aristotle 152 – 154
arithmetic 1, 5, 10, 19, 33, 44, 49, 93, 184
Artificial General Intelligence (AGI) 65, 71, 77
artificial intelligence 1 f., 5 f., 13 f., 21, 25, 27 f., 45, 49 – 53, 56 f., 63 – 68, 70 – 72, 74, 77 – 81, 84 f., 149 – 151, 155, 186, 188
artificial intelligence (AI) 1 f., 5, 13, 21, 27, 45, 64, 186
Asimov, Isaac 3
Askvik, E.O. 184
atomic bomb 7, 64, 78 – 81, 167 f., 188
authentic human creativity (AHC) 51 f., 56 – 58, 69

Babbage, Charles 48 f.
Babbage Engine 48
babbage-in, babbage-out principle, the 49 f.
babbage-in, babbage-out principle, the (BIBO) 48 – 51
babbage's principle argument, the 69
Baby Theresa 99
Baiasu, S. 3
Balzac 59
Bard 2, 85, 188
basic 2 – 12, 19 f., 23, 31, 35, 39, 41, 54 f., 80 f., 84 f., 89, 102, 104, 153 – 156, 158 – 160, 162, 164, 170 f., 179, 181, 183, 188
BBC News 94
Bearzi, M. 97, 101
Bekoff, M. 97
Benacerraf, Paul 10, 41
Bengio, Yoshua 70 f., 80
Benjamin, Walter 58, 62 f.
big-I imagination 56
Blade Runner 24, 83
Blade Runner 8, 23 f., 27, 83, 182 f., 188
BLAH 33 f.
Block, N. 17 f., 28, 66
Bob 24, 29 f., 43, 82
Boolos, G. 4 f., 17, 27, 42, 44, 46
Boonin, D. 98
Bostrom, N. 64, 155
brain-in-a-vat (BIV) 82
brain organoids 13, 82 – 83
Bregman, R. 173
Brian, D. 39
"Brillo Boxes" 61 – 63
Bringsjord, S. 155
broadly Kantian nonideal dignitarian moral theory 2 – 5, 10 f., 14, 20, 23 f., 26 f., 31, 33 f., 40 f., 149, 156, 162, 180 – 182
broadly Kantian nonideal dignitarian theory of moral principles 6, 13, 15, 18, 23
broadly Kantian nonideal dignitarian theory of moral principles, the 23, 28
Bromell, Nick 104 – 106
Bryson, J. 155
Burns, C. 76
Butlerian Jihad 51 f., 57 f., 62

https://doi.org/10.1515/9783111260525-012

Call, J. 1–4, 6–9, 11, 13, 15–17, 20–25, 27, 29, 32f., 35, 39–41, 45, 47, 49–52, 54–56, 64, 66f., 73f., 76–78, 80, 86, 90, 93, 98–102, 104f., 152–157, 159, 163f., 169–172, 175, 177, 182, 186–188
Calvarly, D.J. 155
Camus, A. 33, 38
Cantor, G. 47, 92f.
Cantor's diagonal argument 47
caring 1, 4f., 13, 15, 22, 27, 38, 50f., 55, 80f., 84, 86, 98, 103, 175, 188
Carroll, Lewis 31, 37f., 43
Casablanca 40f.
Categorical Imperative, The 15
caveat lector sentence 32
Chalmers, D. 18, 67
character 26f., 29–31, 35–37, 39–42, 57, 60, 62, 68, 106, 175, 184
chatbots 21, 65, 71, 74, 78
ChatGPT 2f., 20f., 23, 27f., 57, 65, 71, 74f., 79f., 85, 183, 188f.
chinese room argument against strong AI, the 5, 66, 68
Chomsky, N. 2, 39, 69, 81
Clarke, R. 155
Coeckelbergh, M. 155
coercion 107, 155–157, 160, 162, 176, 179–181
computers 1, 6, 13, 42, 63, 80, 83f., 86, 186, 188
conceptual content 5, 20, 37, 39
concrete 5, 17, 34, 40, 63, 75, 167
consciousness 1, 4f., 18, 22, 25, 27, 29f., 38, 41, 51, 58, 65–67, 83f., 92, 98, 103f., 150
constructivism 11f., 18, 26, 34
context-sensitive 28, 32f.
contradiction 2f., 8, 16–20, 26, 28f., 32–34, 41, 176
Corbyn, Z. 189
counterpart thesis, the 17, 28, 66
COVID-19 pandemic 87, 158, 170
Crary, J. 172
creative human rationality for uncomputable functions, the thesis of 47f.
creative materials 58–63
creative piety 55f.
creative rage against the computing machine 51f., 183
creative transformation 50, 54f.
creativity 28, 56, 81, 178
critical philosophy of digital technology, the 52
crypto-Swedenborgian 17, 20
Curtiz, Michael 40f.
cybernetic 1, 3f., 6f., 11, 149, 163, 174, 177–183
cybernetics 1

Dafoe, A. 10
daily dignitarian digital refusards (DDDRs) 188
Dammbeck, L. 72
Danto, A. 61
data or information 5, 157–159
Davis, A. 155, 161
Davis, E. 155, 161
Dawkins, R. 173
De Bruin, L. 164
decidable 11, 17f., 28, 44, 46, 53f., 57f., 63, 66, 165
deductive reasoning 48–49, 53
Deep Consciousness Thesis, The 67, 100
Deep Thought 13
Dennett, Daniel 169, 173
Depeche Mode 1, 36
Deresiewicz, W. 56f.
Descartes's evil demon 83
desires 11, 18, 20, 26, 31, 58, 87, 91, 99f., 103, 175
Dick, Philip K. 83
Dietrich, E. 155
digital computer 2, 5, 11, 14, 16–18, 20–23, 27f., 42, 44–46, 54, 66, 70, 82, 166
digital data or information 1, 5f., 157, 186
digital ethics 3f., 6, 8, 78, 149, 151f., 154f., 160
digital ethics→virtue ethics 151–154
digital technology 1, 3–7, 10–12, 15, 17, 50–53, 57, 59f., 62–69, 77f., 80–83, 85f., 107, 149, 157–159, 162–164, 166–169, 172, 177–179, 182f., 185–189
dignitarian cybernetic instrumentalism 3f., 11f., 85f., 149, 163, 177, 179, 182, 186f.
dignitarian cybernetic instrumentalism (DCI) 3, 11, 85, 149

dignitarian digital ethics 3 f., 11 f., 85, 89 f., 149, 151, 153 – 155, 158, 160, 162 f., 177, 182, 186
dignitarian digital ethics (DDE) 3, 11, 85, 149
dignitarian neo-Luddism with respect to digital technology 4, 52, 80, 163, 177, 186 – 188
dignity 1, 11, 13, 15, 35, 38, 48, 81, 85, 90 – 93, 95 f., 101 f., 104 – 107, 156, 158 f., 161, 169, 174, 176, 179, 181, 183, 189
dignity-skeptic, the 87, 90
Do Androids Dream of Electric Sheep? 24, 82 f.
Donagan, A. 27
Dostoevsky 33, 106
Douglass, Frederick 104 – 106
Dowe, D. 22
Dreyfus, Hubert 66, 68
Dr Frankenscience of digital technology and the atomic bomb, the 72 f., 166 f., 169
dual-content cognitive semantics 39
Duchamp, Marcel 61, 63
Dune 51 f.
duties 1, 6 – 12, 19 – 24, 26, 29, 36, 101
Dyson, F. 44, 72, 166

Economic and Philosophical Manuscripts 95
Ehrenreich, B. 184 f.
Einstein, Albert 31, 39, 43
Einstein syndrome 39
Eisenhower, President Dwight D. 2, 173
Ekirch, A.R. 172
Else, J. 1, 27, 30, 33 f., 39, 43, 46, 57, 72, 93 f., 106, 156, 158
Emerson, R.W. 1, 4, 28, 41
enactive-transformative principle, the 72
enantiomorphic 29 f., 68
enlightenment 163, 174 f., 177 – 179, 183
Essential Embodiment Theory, The 5, 24, 67
essentially embodied 4, 13, 15 f., 18 f., 24, 27 – 29, 51, 55, 58, 60 – 62, 64, 67, 84, 87, 89 – 91, 95 – 97, 103 f., 164 f., 175, 183 f., 188
essentially embodied selves thesis, the 16 – 19
essentially non-conceptual 29, 37, 41, 66, 81, 95, 165
essentially non-conceptual content 20, 37, 39
ethics 2, 4, 6, 11, 13, 26 f., 34 f., 38, 51, 55, 149 – 151, 153, 162, 174 f.
Etinson, A. 87
Excluded Middle Constraint, The 26, 30 – 32, 34, 36 f.
existential 3, 12, 20, 26 – 28, 38, 41, 55, 63, 65, 70 f., 78, 80 f., 107, 151 f., 188
explosion, logical 16 – 18, 28 f., 33 f., 41, 167

faculty of desire 1 – 3, 5 – 9, 12 – 16, 18 – 21, 23, 27 – 32, 34 f., 41 f., 45 f., 48, 51, 56, 60 – 62, 64 f., 71, 73 – 79, 82 – 86, 93 f., 97, 99 f., 103, 105, 150 f., 167, 169, 175, 179 f., 183, 188 f.
fairly universal and objective moral principles 6 – 8, 11 f., 19 f., 22 – 26, 29 – 32, 36
Fallacy of Inevitability, The 6 f., 10
Farhany, N. 189
Feynman, Richard 166 f.
Field, D.G. 9, 23, 82, 155, 167, 183
Fisher, M. 178
Flatlander Fallacy, The 33, 36, 40
Floridi, Luciano 63 f., 155
flourishing 21 f., 24 – 26, 152, 155 – 157
Fong, Benjamin Y. 178
“Fountain” 61 – 63
France 2
Frankenscience 72 f., 166
Frankfurt, Harry 93, 99 f., 102
Frankfurtian human real person (person$_f$) 103
free agency 1, 4 f., 11, 16, 20, 22, 27, 38, 51, 55, 57, 68, 84, 97, 99, 103, 157
freedom 14, 20, 57, 95, 100, 103, 106, 158, 175 f., 178, 182, 185
Frege 5
French, P. 2, 35, 40, 160
Fromm, E. 95
functionalism 17 f., 70
functions 5, 9, 39, 44 – 48, 63, 68 f.
Future of Life Institute 77 f.

Gallagher, S. 164
garbage-in, garbage out principle, the 48 f., 69
garbage-in, garbage out principle, the (GIGO) 48 – 49, 69
Gare, Arran. 16, 20
Geist/ghost-self or spirit 17, 19
Generative Pre-trained Transformer 21, 24 – 26, 75 – 78

Generative Pre-trained Transformer (GPT) 21, 24
genius 55, 106, 169
"German cannibals" case 94
Gips, J. 155
Glendinning, C. 4, 186
gmail 23, 27
God 22, 92, 169
Gödel, Kurt 16, 19, 33, 44, 47, 49, 53 f., 69
Gödel's incompleteness theorems 19
Godwin, William 179
Goldstine, H. 166
Google 24, 65, 70, 76
Gordon, J.-S. 155
Graeber, D. 170
Grau, C. 154 f.
Griffin, D.R. 97, 101
Griffiths, M.D. 178
Guarini, M. 155
Gunkel, D.J. 155
guns 7, 10

Hall, J.S. 155
halting problem, the 42, 46, 68
Hampson, R. 155
Hanna, R. 2, 4 – 6, 8 – 12, 15 f., 18 – 20, 22, 24, 27, 29 – 31, 33, 38 – 41, 43, 45, 47, 51, 53, 55 f., 62 f., 67 f., 70, 72, 80 – 82, 84, 89 – 91, 93 – 95, 98 – 101, 103 f., 106, 154, 156, 164 f., 170 – 175, 177, 179, 183, 185
Harris, S. 173
Hawkins, J.M. 1, 33, 35, 97, 186
heideggerian argument against strong AI, the 66, 68
Herbert, Frank 51 f., 183
Hierarchy of Kantian Moral Principles, The 23 – 25
highest good, the 1, 93, 97
Hill, T. 4, 11, 27
Hinton, Geoffrey 70 f., 77 – 80
Hitchhiker's Guide to the Galaxy, The 13
Horn, M.B. 1, 3
"human, all-too-human" 3, 8, 16 f., 20, 22, 32, 34, 40, 71, 85, 91, 106, 159, 169
human dignity 1 – 4, 7 f., 10 f., 14, 20, 38 f., 73, 80 f., 84 – 87, 89 – 97, 104 – 107, 149, 155 – 157, 159 – 161, 167, 178 f., 181 f., 187 – 189
human real personhood 22, 56, 80 f., 84 f., 87, 90, 95 f., 100, 102 f., 156, 160, 188
human real persons 1, 14, 19, 21, 23, 83, 85, 87, 89 – 93, 95 – 102, 155 – 157, 159 – 162, 174
Hume, D. 102

illegibility 32
illegible 32, 37, 43
imagination 56
Imitation Game, The 20 – 23, 27 f., 66, 68, 82
inadequacy of the turing test argument, the 68 f.
incompleteness argument, the 69
indexical 25, 28 f., 31 – 33, 36, 66
individual S-type animal 98 f.
inductive reasoning 53
inherently repeatable objects 41
innate faculties 104
innate powers 104
intellectual cognition 4 f., 22, 27, 38, 51, 84
intelligence 7, 14, 17 – 19, 21 – 23, 27 f., 41, 51, 63, 65 f., 68, 77, 80 – 82, 84 f., 149 f., 188
intentionality 4 f., 22, 27, 38, 61, 66 f., 98, 103
intuition 2, 8 – 10, 18 – 20, 23, 41, 81, 89, 96, 167 – 169
Invasion of the Body Snatchers 86
invasion of the mind snatchers, the 65, 86, 188
it's-all-done-with-mirrors argument, the 68 f.

Jabberwocky 31, 37, 43
Jackson, C.J. 1, 3, 18
Jauhar, S. 38
Jeffrey, R. 4 f., 17, 27, 42, 44, 46

Kaczynski, Theodore 182
Kant, I. 1, 3 – 8, 12 – 16, 18, 23 f., 26 f., 29 – 31, 33 – 40, 47, 55, 86, 90, 99, 101, 105 f., 174, 183
kantian human real person ($person_K$) 102 f., 159, 161 f.
Kantian modal dualism 4 f., 47
Kant-Sartre Insight, The 3, 11, 38
Kant the Foundations of Analytic Philosophy
Keller, A. 19, 45, 69, 81, 84
Kim, J. 17, 28, 66
kindness 21 f., 94

King Jr, Martin Luther (MLK) 106, 180 f.
King Lear 39
Kleingeld, P. 106
Körner, Edith 110, 139
Körner, Stephan 110, 139
Korsgaard, C. 4, 27
Kranzberg, M. 170 f.
Kripkenstein 45 f.
Kripke, Saul 5, 33, 45, 47
Kropotkin, Peter 189
Kuhse, H. 97
Kurzweil, R. 64

Laker, B. 158, 160
LaMDA 80, 85, 188,
Landgrebe, J. 19, 45, 68, 81, 84
Large Language Models 2, 24 f., 57, 65, 71, 74, 76, 78 – 80, 82 f., 85 f., 188
late-talking syndrome 39
Lector Sentence, The 32 – 36
legibility 32, 37
legible 32, 37
Lemoine, Blake 23
Lesser Evil Principle, The 11, 29 – 32, 34, 36, 39 f.
Lesser Evil Principle*, The 40
Lewin, R. 101
Liar Sentence, The 16, 32 f., 35
lies 35, 37 f., 174
Life-Shaping Thesis, The 164 – 166, 168
logging off 178
logic 4 f., 7, 15 – 19, 28, 31 – 33, 39, 44 f., 47, 49, 59, 63, 184
Loh, J. 155
Lovely, Garrison 44, 65, 71 f.
Luddism 4, 182, 186
Luddite 4, 80, 186 – 188

Mackie, John 9
Mackworth, A.K. 155
Maiese, Michelle 5, 11, 16, 18, 24, 27, 29 f., 51, 55, 62 f., 67 f., 70, 72, 80, 84, 89, 91, 94 f., 100, 156, 164 f., 172 f., 175
manifestly real 4, 16 f., 20, 40, 42, 47, 54, 60, 62, 92
Marcus, G. 155
Marcus, Ruth Barcan 140
Marx, Karl 95
Mason, H.E. 4
master 1, 11
McDermott, D. 155
McDonald, A. 7
McGinn, C. 67
McKeever, M. 166
McLaren, B.M. 155
McWhorter, L. 161
mechanistic worldview, the 4, 11, 15, 70, 165
Mellin, William D. 48
metaphysical ground 80, 84, 96, 188
metaphysics 1, 3, 5, 15 f., 19, 85, 87, 91, 106, 175
Metz, C. 79
military-industrial complex, the 2, 173
military-industrial-digital complex, the 2, 51, 58, 70 – 72, 77, 81, 163, 173, 177 – 180, 182, 188 f.
Miller Jr, W.B. 11, 53, 67, 70
Mill, J.S. 182
mind 1, 3, 5, 10 – 12, 17 – 20, 22, 24, 28, 47, 51, 56, 70, 77, 82, 86, 91, 97, 149 f., 164 f., 169, 172 – 178
mind-body politic, the 72, 164 f.
minimal principle of excluded middle, the 34
minimal principle of non-contradiction, the 19, 33 f., 81
mirrors 28, 168, 184
Misselhorn, C. 155
moist robots 169, 173
monotonic 53
Moor, J.H. 155
moral 1 – 16, 18 – 41, 49, 52, 55, 58, 81, 84 – 87, 89, 92 f., 95 f., 101 – 103, 105 f., 149 – 163, 166 f., 174 f., 179, 181 f., 186 f.
Moral Doomsday Scenario, The 27 f., 32 – 34, 41
morality 1, 5 f., 10, 14 – 16, 26 f., 33, 38, 51, 55, 92, 99, 105, 152 – 154, 160 – 162, 175
Muldoon, J. 3
Müller, V.C. 155
Musk, Elon 77 f.
myth of AI 51, 58, 64
myth of artificial intelligence, the 51, 63 – 66, 69 – 72, 80, 82

naturalistic fallacy, the 7
Nazis 3, 32, 37, 40, 94, 166
neo-Luddism 186
neo-organicist worldview, the 4, 11, 13, 165
Newen, A. 164
Nexus VI replicants 83
Nguyen, C.T. 178
Nietzsche 16, 22
nocebo 173
No-Foolish-Consistency Interpretation, The 4f., 29, 32–36, 41
No-Global-Violation Constraint, The 24f., 31f., 34, 36
non-authoritarian 2, 7, 20, 52, 158, 161, 172, 179f., 182
non-denumerable 44f., 47, 92f.
non-monotonic 53
nonviolent 157, 179–181
nuclear weapons 72f., 76, 78–81, 166
numbers 10, 44, 47, 92f.
Nussbaum, M. 26
Nyholm, S. 155

Oakes, K. 28
O'Connell, M. 14
old-school, phenomenological critique of strong AI, the 30, 66, 69
O'Neill, O. 4, 11
OpenAI 2f., 21, 23f., 27, 71, 73–77
Oppenheimer, J. Robert 72f., 79
oppression 1, 4, 11, 51, 95, 107, 149, 157f., 163, 177, 179, 181, 183, 189
Oppy, G. 22
Optical Character Recognition 41f., 47
optical character recognition (OCR) 41, 47
optimization 6
organicity argument, the 68f.
organoid intelligence (OI) 13, 82, 86, 188
organoids 13, 82–84
Ostrom, E. 171
ought implies can 7, 32
Oxford Encyclopedic English Dictionary 1, 33, 35
Oxford English Dictionary 35, 97
oxymoron 2, 5, 64

Paans, Otto 4, 11f., 39, 53, 55, 67, 70, 82, 164f.
Page, Larry 2, 23, 65, 70, 78f.
paradox 3, 28, 33
Paris 2, 61
Parshley, L. 65, 187
Parsons, Charles 10
Pascal, B. 166, 172
Peano axioms 19, 49
Peer-to-Peer Foundation, The (P2P) 171
perceptibility condition, the 35–37, 43
Philosophical Investigations 30–32, 39, 43, 45f., 96
Pittenger, Mark (MP) 58
placebo 173
Plato 154
Popova, M. 14, 16
posthumanism 14f., 155
Powers, T.M. 1, 16f., 29, 51f., 55–57, 64f., 71, 104f., 151
practical reason 6, 13, 18, 21, 24, 26
Principia Mathematica 19, 33
principled authenticity 102–104
principle of minimal legibility, the 33f.
privacy 158–160
problem of digital technology, the 2f.
Professor Von Braun 2
Putnam, Hilary 1, 4–7, 14, 16–18, 31, 33, 35, 37, 39, 41, 47, 68, 72, 80–84, 87, 90, 94f., 101f., 159f., 163, 166, 174, 176, 187

Quine, W.V.O. 5, 18

Rachels, J. 99, 153f.
Rachels, S. 99, 153f.
Rage Against the Machine 52
Ran 39
Rapaport, M. 60
rational consent 87, 94f., 157, 159f.
rational human 2, 4–6, 8f., 11, 13, 15, 17–20, 22–24, 27–31, 34, 36, 38–43, 45–48, 50f., 56, 62–70, 80f., 83f., 87, 91, 102, 154, 164, 170, 174, 184, 188
rational intuition 8f., 27
rationality 2, 15, 22, 33, 46, 48, 89, 99–103, 154, 175
Rawls, John 9, 159
readability 31f., 34f., 37–41
readability argument, the 68f.

readable 30–37, 41–43, 60, 62
reading 4, 14, 29–33, 35, 37–43, 47, 174, 184
real human personhood 98, 103
Realm of Ends, The 1, 13–15, 20, 37, 102
responsibility 3, 38f., 55, 57, 101, 103, 106, 161
revolution of the will 105
Right and the Good, The 1, 4, 8–10, 15, 36
Roberts, I. 2, 69, 81
Robobob 29f., 43, 82
robot 1–62, 64–73, 76, 78–87, 89–107, 149, 151–166, 168–184, 186–189
Israel, J. 175
robotics 1, 5f., 67, 186
Royce, Josiah 89
rule 1, 5f., 8, 11, 17f., 29, 31f., 39, 45–47, 66, 68, 154, 160, 162, 164, 168
rule-following paradox, the 45f.
rule-governed 6, 44f.
Russell, Bertrand. 5, 37

Sartre, J-P. 2f., 26, 40f.
Satan 22
Savage-Rumbaugh, S. 101
Sayre-McCord, J. 27
Schapiro, T. 30
Schopenhauer, A. 169
Schrödinger, E. 47
Schrödinger's cat 47
Schuessler, J. 169, 173
Schulson, M. 178
Scott 8, 23f., 27, 182f., 188
Scott, J.C. 8, 23f., 27, 182f., 188
Scott, Ridley 24, 83
Searle, John 5, 66
self-consciousness 4f., 22, 27, 38, 51, 83f., 92, 101, 105f.
self-referring 32
semantic condition, the 35, 37f., 43
sensible cognition 4f., 22, 27, 38, 51, 84
Seville, H. 155
Shadbolt, N. 155
Shafer-Landau, R. 152, 154
Shapiro, Stewart 10
Shelley, Mary 72f.
Shiffrin, S. 27
Siebert, C. 101
Siegel, Don 86
Singer, P. 97
Singularity 11, 64
Sipser, M. 17
Sisyphus 38, 169
slavery 7f., 10f., 161, 171f., 174, 177f.
Sloane, Mona 8–10
Smirnova, L. 82, 84f.
Smith, B. 19, 45, 68, 81, 84
Smith-Garcia, D. 39
Smit, I. 155
social contexts 62f.
socialism 175, 178
Sophie's Choice 3, 32
sortal correctness 37
soundness vs. validity principle, the 49
Sowell, T. 40
Spinoza 175
spiritualism 14, 16f., 20, 70
spiritual selves 16, 19f.
Stanford, C. 21, 97, 101
Statism 179
strong thesis of artificial intelligence, the 14f., 17–19, 28–31, 40, 43, 66–69, 73, 82, 155, 173
strong thesis of artificial intelligence, the (strong AI) 15, 17, 28, 31, 66
Sullins, J.P. 155
Superalignment Generalization Team, The 73–76
superalignment problem, the 73f., 76
superintelligence 64, 74
supra-heapists 59
Swedenborg 15
Swedenborgian spiritualism 15
Sydney 2, 85, 188
Sympathy 102, 167
syntactic condition, the 35–37, 43
synthetic a priori 15f., 18f., 27, 80f.

Tarski, A. 5, 16, 33
technological determinism 10
texts 2, 9, 11f., 19, 22, 33–36, 40–42, 49, 54, 59–61, 63, 65, 69, 105, 151–153
"The Algorithmic Auditing Trap" 10
The Kant Wars 3–5, 7, 11–16, 19–21, 26f., 31, 34–36, 39–41, 45, 55, 89, 91, 93, 99, 104–106, 151, 170, 174f., 177, 179

thesis of formal mechanism, the 17, 28, 66
thoroughly nonideal world 3, 7, 34
thought-shapers 39, 164–166
toddlers 101f.
Tomasello, M. 101
tool 1, 3, 6, 9–11, 26, 57, 60, 62, 85, 149, 164, 168
Torday, J. 11, 53, 67, 70
Torrance, S. 4, 149, 151, 155
transcendental 4, 20, 47, 93, 97
transfinite 44f.
transhumanism 14
true contradictions 16
truth-value gluts 16f.
Turilli, M. 155
Turing, Alan 2, 5, 11, 14, 16–18, 20–23, 27f., 42, 44–46, 48, 54, 63, 66, 70, 82, 84, 166, 173
Turing-computation 5f.
Turing machine 17–19, 21f., 27–29, 31, 42–45, 48, 59f., 62–64, 66, 84, 173
Turing Test, The 21f.
Turkle, Sherry 167f.
Two Slit Experiment, The 47

Ultimate Question, the 13
ultimate sourcehood 55, 57
Unabomber, The 72f.
Unamuno, M. 33
uncomputable 19, 33, 44, 46f., 63
uncomputable functions argument, the 68f.
undecidable 11, 18f., 44, 46–48, 53, 165
unprovable sentence 16, 47
unreadability 32, 36
unreadable 32–37, 42f., 68
US Air Force killer-drone operators 168
utilitarianism 154

Vallor, Shannon 151f.
van der Meer, Audrey 184
van der Weel, F.R. 184
violent 52, 58, 157, 180–182
virtual reality 83, 167
virtue ethics 151–154
Volition 4–6, 11, 13f., 18, 20, 22, 24, 27, 38, 51, 84, 100, 103

Wachowski, Lana 41f., 83
Wachowski, Lilly 41f., 83
Wallach, W. 155
Warhol, Andy 61, 63
weak AI thesis, the (weak AI) 18f.
Weber, M. 179
Whitby, B. 155
Wikipedia 2f., 15, 21, 23, 52, 57, 168
will 6, 18, 24
Williams, B. 26, 154
Wittgenstein, L. 43, 96
Wizard of Oz, The 22
Wozniak, Steve 78

Yetish, G. 172
Yudkowsky, E. 155

Zermelo, E. 92